Two Applications of Wild Bootstrap Methods to Improve Inference in Cluster-IV Models

Keith Finlay[*] Leandro M. Magnusson [†‡]

US Census Bureau University of Western Australia

November 7, 2018

Abstract

Microeconomic data often have within-cluster dependence. This dependence affects standard error estimation and inference. When the number of clusters is small, asymptotic tests can be severely oversized. Additionally, in the instrumental variables model, the potential presence of weak instruments introduces an additional caveat for carrying out hypothesis testing. We study the use of wild bootstrap methods to improve inference in two empirical applications that rely on IV models with cluster dependence. Building from estimating equations and residual bootstrap methods, we identify variants that are robust to the presence of weak instruments and a small number of clusters. These methods reduce absolute size bias significantly and demonstrate that the wild bootstrap should be considered as part of the standard toolkit in IV and cluster-dependent models.

Keywords: instrumental variable, hypothesis testing, weak instrument, clustered error, wild bootstrap.

JEL codes: C12, C15, C31.

[*] US Census Bureau, Room 5K132D, 4600 Silver Hill Road, Washington, DC 20233, USA, kfinlay@gmail.com.

[†] Department of Economics, 35 Stirling Highway M251, Business School, University of Western Australia, Crawley, WA 6009, Australia, phone: +61 (8) 6488-2924, fax: +61 (8) 6488-1016, leandro.magnusson@uwa.edu.au (corresponding author).

[‡] We would like to thank David Hendry, Jan Ondrich, Maurice Bun, Colin Cameron, and participants at the 2012 Econometric Society Australasian Meeting and at the 2015 International Association for Applied Econometrics Conference. We also thank the National Science Foundation (award SMA-1004569) for support. The results and conclusions in this paper are the personal views of the authors and do not necessarily reflect the views of the US Census Bureau.

1 Introduction

Microeconometric data often have a group structure. When regression errors are correlated within these groups or clusters, it is well-known that variance estimates can be biased and that hypothesis testing can be misleading. The common solution to this problem is to use cluster-robust standard error estimation methods whose asymptotic properties rely on a large number of clusters. When the number of clusters is small, the rejection rates of tests can be well above their nominal levels even when cluster-robust standard errors are used (see Cameron et al. (2008)).

In the linear instrumental variables (IV) model, statistical inference is further complicated by the possibility of weak instruments. Several tests have the correct size asymptotically when the instruments are weak, such as the cluster-robust versions of the Anderson and Rubin (1949) (AR), Kleibergen's (2002) Lagrange multiplier (KLM), and Moreira's (2003) conditional likelihood ratio (CLR) tests.[1] We show in our Monte Carlo experiments, however, that these tests experience the same size distortions as the Wald test does when the number of clusters is small, even when instruments are strong.

Bootstrap methods can improve the reliability of inference when sample sizes are small. The works of Cameron et al. (2008), Kline and Santos (2012), and MacKinnon and Webb (2016) highlight the use of bootstrapping to improve inference when there is intra-cluster dependence in the linear model with only exogenous covariates. They show that a variant of Wu's (1986) wild bootstrap method with cluster-based sampling performs well in a variety of cases, and bootstrap tests dominate the asymptotic tests in terms of size. It is well-known, however, that bootstrapping cannot improve the performance of the Wald test when instruments are weak (Davidson and MacKinnon, 2008; Moreira et al., 2009; Zhan, 2014).

In this study, we focus on two problems. The first is that, with few clusters, the asymptotic critical values for the tests can be poor approximations of the corrected finite sample critical values. The second is the potential presence of weak instruments, which make inference unreliable even for a large number of clusters. Therefore, we propose wild bootstrap methods for the AR test that can make inference more reliable when the number of clusters is small and instruments are potentially weak.

We use two applications from prominent political economy studies to explore the importance of accounting for few clusters in the IV model: the effect of institutions on economic growth discussed by Acemoglu et al. (2012) and the impact of the economy on civil conflict discussed

[1] Other statistical tests have asymptotic and nominal size equality independent of the presence of weak instruments, such as the conditional linear combination test of Andrews (2016).

by Miguel et al. (2004). In these applications, the specifications vary in terms of the number of clusters, the instrument strength, and the controls. We find large differences between the confidence sets derived from the asymptotic and bootstrap tests, even in cases where the effective F-test of Olea and Pflueger (2013) is above the 5% critical value. Bootstrapping, therefore, is consequential for inference in specifications that add controls, stratify the sample, reduce sample variation, or reduce the number of clusters. We recommend that our wild bootstrap tests be used as diagnostic tools and to conduct inference in situations where there is any uncertainty about the effect of cluster sampling or instrument weakness.

We perform extensive Monte Carlo simulation experiments to explore the performance of our bootstrap methods. We consider data generating processes (DGP) with various error distributions, increasing heterogeneity among errors, and different cluster sizes and instruments. We find rejection rate levels as high as 50% with the cluster-robust Wald test when the nominal level is 5% in a strong-instrument scenario with 20 clusters. With our cluster *estimating equations* and *residuals* bootstraps, we obtain rejection rates that are very close to the nominal levels in the same experiments.

Our methods tackle a number of challenges examined in isolation in previous studies. In a previous study, Gelbach et al. (2007) propose a variant of the wild cluster bootstrap method of Cameron et al. (2008) for the Wald test in an IV setting that is valid only if the instruments are strong. Davidson and MacKinnon (2008) propose several bootstrap techniques for linear IV models assuming that the residuals are homoskedastic, and they further extend these techniques by allowing residual heteroskedasticity, but only at the individual level (Davidson and MacKinnon, 2010). Using Edgeworth expansions, Kleibergen (2011) shows that the bootstrap method decreases the size distortion of the AR test.

In the next section, we establish a minimal model to frame the discussion of inference in a cluster-sample IV model. That section is followed by the empirical applications of the proposed bootstraps. More detailed descriptions of the tests and their bootstrap counterparts are in the fourth and fifth sections, respectively. In the sixth section, we present the simulation design and Monte Carlo experiment results. Several bootstrap methods, including the ones for the KLM and CLR tests, and further simulation evidence are included in the supplement that accompanies this manuscript.

2 Cluster-robust inference in a simple IV model

A simple linear IV model of cluster-sample data with G clusters is

$$\begin{cases} \mathbf{y}_{1,g} = & \mathbf{y}_{2,g}\theta + \mathbf{u}_g \\ \mathbf{y}_{2,g} = & \mathbf{z}_g\Pi_z + \mathbf{v}_g \end{cases} \quad \text{for } g = 1, \ldots, G, \tag{1}$$

where $\mathbf{y}_{1,g}$, $\mathbf{y}_{2,g}$, $\mathbf{z}_g$, $\mathbf{u}_g$, and $\mathbf{v}_g$ are $n_g \times 1$ vectors, and n_g is the number of observations in cluster g. The total sample size is $n = \sum_{g=1}^{G} n_g$. We assume that the errors have an arbitrary covariance structure within clusters but are independent across clusters. Without loss of generality, we exclude exogenous regressors, multiple endogenous variables, and multiple instrumental variables. In this problem, the IV estimator is $\hat{\theta}_{\text{IV}} = (\mathbf{Z}'\mathbf{y}_2)^{-1}\mathbf{Z}'\mathbf{y}_1$, where $\mathbf{Z} = [\mathbf{z}_1', \ldots, \mathbf{z}_G]'$, $\mathbf{y}_2 = \left[\mathbf{y}_{2,1}', \ldots, \mathbf{y}_{2,G}'\right]'$, and $\mathbf{y}_1 = \left[\mathbf{y}_{1,1}', \ldots, \mathbf{y}_{1,G}'\right]'$ are $n \times 1$ vectors.

When the errors are assumed to be independent and identically distributed (iid), the estimator of the variance is $\widehat{\text{Var}}_h(\hat{\theta}_{\text{IV}}) = \hat{\sigma}_u^2 (\mathbf{y}_2' \text{P}_{\mathbf{Z}} \mathbf{y}_2)^{-1}$, where $\hat{\sigma}_u^2 = \frac{1}{n}\hat{\mathbf{u}}(\hat{\theta}_{\text{IV}})'\hat{\mathbf{u}}(\hat{\theta}_{\text{IV}})$ and $\hat{\mathbf{u}}(\hat{\theta}_{\text{IV}}) = (\mathbf{y}_1 - \mathbf{y}_2\hat{\theta}_{\text{IV}})$.[2] However, in the presence of intra-cluster dependence, even if this dependence is negligible, we can use the same arguments as in Moulton (1990) to show that $\widehat{\text{Var}}_h(\hat{\theta}_{\text{IV}})$ underestimates the variance of $\hat{\theta}_{\text{IV}}$. The most commonly used estimator of $\text{Var}(\hat{\theta}_{\text{IV}})$ is $\widehat{\text{Var}}\left(\hat{\theta}_{\text{IV}}\right) = (\mathbf{Z}'\mathbf{y}_2)^{-1}\left[\sum_{g=1}^{G}\mathbf{z}_g'\hat{\mathbf{u}}_g(\hat{\theta}_{\text{IV}})\hat{\mathbf{u}}_g(\hat{\theta}_{\text{IV}})'\mathbf{z}_g\right](\mathbf{Z}'\mathbf{y}_2)^{-1}$ where $\hat{\mathbf{u}}_g(\hat{\theta}_{\text{IV}}) = \mathbf{y}_{1,g} - \mathbf{y}_{2,g}\hat{\theta}_{\text{IV}}$ is the vector of IV residuals for the g^{th} cluster. This estimator is an adaptation of the Huber-White heteroskedastic-robust sandwich estimator (White, 1980; Arellano, 1987), which does not impose any structure on the variance of the error term.

We are interested in making inference about θ. For example, we may want to test the following hypotheses: $H_0^\theta : \theta = \theta_0$ against $H_1^\theta : \theta \neq \theta_0$. The Wald test for the structural parameter is

$$W(\theta_0) = \frac{\left(\hat{\theta}_{\text{IV}} - \theta_0\right)^2}{\widehat{\text{Var}}\left(\hat{\theta}_{\text{IV}}\right)} \xrightarrow{d} \chi^2(1),$$

where $\xrightarrow{d}$ indicates convergence in distribution as $G \to \infty$ and $\chi^2(1)$ is the χ^2-distribution with one degree of freedom. The null assumption is rejected if $W(\theta_0) > \chi^2_{1,1-\alpha}$, where $\chi^2_{1,1-\alpha}$ is the $1-\alpha$ quantile of the $\chi^2(1)$-distribution.

As the instrument $\mathbf{z}_g$ becomes weakly correlated with the endogenous variable $\mathbf{y}_{2,g}$ ($\Pi_z \to 0$), the IV estimator $\hat{\theta}_{\text{IV}}$ becomes inconsistent. Consequently, the Wald test does not have the correct

[2] We use the notations P_A and M_A for the projection matrices $\text{P}_\text{A} = \text{A}(\text{A}'\text{A})^{-1}\text{A}'$ and $\text{M}_\text{A} = \text{I} - \text{P}_\text{A}$ throughout the text.

asymptotically size (Staiger and Stock, 1997).[3] Olea and Pflueger (2013) propose the effective F test for testing instrument weakness when errors are clustered, but only for the case of one endogenous variable. [4]

Instead of pre-testing instrument strength before computing the Wald test, we can simply use a test that is asymptotically valid independent of whether the instrument is strong or not. Consider now the following representation of model 1:

$$\begin{cases} \mathbf{Y}_g\left(\theta_0\right) = & \mathbf{z}_g \delta_z\left(\theta_0\right) + \mathbf{e}_g\left(\theta_0\right) \\ \mathbf{y}_{2,g} = & \mathbf{z}_g \Pi_z + \mathbf{v}_g \end{cases} \quad \text{for } g = 1, \ldots, G, \tag{2}$$

where $\mathbf{Y}_g\left(\theta_0\right) = \mathbf{y}_{1,g} - \mathbf{y}_{2,g}\theta_0$, $\delta_z\left(\theta_0\right) = \left(\theta - \theta_0\right)$, and $\mathbf{e}_g\left(\theta_0\right) = \mathbf{u}_g + \mathbf{v}_g\delta_z\left(\theta_0\right)$. The parameter $\delta_z\left(\theta_0\right)$ is consistently estimated by the ordinary least squares (OLS) estimator $\hat{\delta}_z\left(\theta_0\right) = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}\left(\theta_0\right)$, independent of the value of Π_z.

We can test $H_0^\theta : \theta = \theta_0$ indirectly by testing $H_0^\delta : \delta_z\left(\theta_0\right) = 0$. This test, the cluster-robust AR test, is defined as

$$AR(\theta_0) = \frac{\left[\hat{\delta}_z\left(\theta_0\right)\right]^2}{\widehat{\mathrm{Var}}\left(\hat{\delta}_z\left(\theta_0\right)\right)} \overset{d}{\to} \chi^2(1).$$

where $\widehat{\mathrm{Var}}\left(\hat{\delta}_z\left(\theta_0\right)\right) = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left[\sum_{g=1}^{G}\mathbf{z}_g'\hat{\mathbf{e}}_g(\theta_0)\hat{\mathbf{e}}_g(\theta_0)'\mathbf{z}_g\right]\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}$ is the variance estimator of $\hat{\delta}_z\left(\theta_0\right)$ and $\hat{\mathbf{e}}_g(\theta_0) = \mathbf{Y}_g\left(\theta_0\right) - \mathbf{z}_g\hat{\delta}_z\left(\theta_0\right)$. The AR test can serve as an alternative to the Wald test, and it has the correct size regardless of instrument strength.

The distributions of statistical tests based on the cluster-robust variance estimator, however, can differ considerably from their asymptotic distributions. Young (2016) and Imbens and Kolesár (2016) show that the cluster-robust estimator of the variance is downward biased in the linear regression with just exogenous covariates, which is the case of $\widehat{\mathrm{Var}}\left(\hat{\delta}_z\left(\theta_0\right)\right)$. Consequently, cluster-robust statistical tests overreject the null hypothesis when it is true. Those studies propose corrections that "inflate" the variance estimator. Our simulation experiments in Subsection 6.2 also show that the cluster-robust Wald and AR test rejection rates are well above the nominal levels under the true null assumption. Instead of correcting the variance, we simply bootstrap the statistical tests to obtain a better approximation of the critical value.

[3] These authors also suggest that a first-stage F-test below 10 indicates weak instruments when the errors are iid. This value has become a commonly used rule-of-thumb by practitioners. However, Bun and de Haan (2010) show that, with a nonscalar error covariance structure, the use of the rule-of-thumb for the standard and the cluster-robust first-stage F-tests combined is a poor guide for detecting instrument strength. Andrews et al. (2018) survey manuscripts published at the *American Economic Review* from 2014 to 2018, and find clear evidence that authors and journals favor specifications that satisfy this rule-of-thumb even when residuals are heteroskedastic.

[4] Stock and Yogo (2005) and Sanderson and Windmeijer (2016) propose similar tests under the assumption that the residuals are homoskedastic.

In the following section, we illustrate the differences between the confidence sets obtained from the asymptotic and bootstrap cluster-robust tests using the applications of Acemoglu et al. (2001) and Miguel et al. (2004).

3 Applications

In the applications that follow, we explore published studies that (1) use IV models, (2) incorporate cluster-sample data with a small number of clusters, and (3) have a risk of weak instruments. We estimate 95% confidence intervals and regions for the structural parameters that are derived from inverting the cluster-robust Wald (Wald Asymp.), Wald multi-equation efficient bootstrap (Wald ME-eff), cluster-robust AR (AR Asymp.), and AR single-equation efficient bootstrap (AR SE-eff) tests. The 95% confidence intervals and sets are formed by the points in the parameter space that do not reject the null hypothesis of being the true parameter at a significance level of 5%. The bootstrap tests impose the null hypothesis for the bootstrap DGP and use Rademacher weights for sampling the residuals. The AR SE-eff and the Wald ME-eff bootstraps are described, respectively, in Sections 5 and S.5 in the Supplement.

3.1 Application 1: The Colonial Origins of Comparative Development

Our first application comes from Acemoglu et al. (2001). In that study, the specifications have only one endogenous explanatory variable and one excluded instrument, with a maximum of 36 clusters.

Acemoglu et al. (2001) study how institutions, as measured by an appropriation risk score, affect economic performance. Identifying the effect of institutions is confounded by its simultaneity with growth and by omitted variables that may affect both variables. The authors argue that the mortality rates faced by Europeans affected their willingness to establish settlements and their choice of colonization strategy. Places where mortality rates are high are likely to have extractive institutions, whereas healthy places are prone to have better economic and political institutions. Therefore, the settler mortality rate would be a good instrument for the institutions variable.

Information about mortality rates is based on historical settler mortality measures, and aggregated regional measures are allocated to match modern countries. In this exercise, one mortality rate can be assigned to several countries (e.g., Latin American and West African countries). Therefore, countries that share the same settler mortality rate constitute a cluster and, consequently, have the same common instrument.

We choose to replicate some specifications from the Acemoglu et al. (2012) reply to the Albouy (2012) comment. The exchange is relevant here because one of the issues of contention is the ability of the instrument to identify the effect of institutions on economic growth. Both studies use a cluster-robust AR statistic.[5]

We work with mortality rates capped at 250 per 1,000 per annum, because several specifications with uncapped rates result in unbounded confidence intervals for both the asymptotic and bootstrap tests. We also focus on specifications that contain a correction for the mortality rate, which is one major criticism of the results of Acemoglu et al. (2012). The correction consists of the inclusion of a dummy "campaign" variable as another control. Albouy (2012) argues that the mortality rates during peacetime and "campaign" episodes are not the same. This correction reduces sample variation, affecting the instrument's identification power, giving us a broad set of practical situations, and demonstrating how instrument strength and cluster structure interact to affect inference.

The results are in Table 1. The first column contains the baseline specifications with no correction for mortality rates. The second column has the same specifications with the correction suggested by Albouy (2012). The third and fourth columns show the results with the minimal and extended corrections of the campaign dummy proposed in Acemoglu et al. (2012). Along the rows of Table 1, the specifications vary according to different sets of regression controls and samples. The same table also reports the effective F-test, which is the same as the cluster-robust first-stage F-test when only one instrument is present and illustrates the variation in instrument strength across specifications. The critical values of the effective F-test are included in the table's notes. Compared with Column (1), the instrument strength measured by the effective F-test drops considerably when controlling for Albouy's "campaign" dummy and remains the same with the extensive correction proposed by Acemoglu et al..

[Table 1 about here.]

We first note that the AR asymptotic confidence intervals are larger than the Wald asymptotic confidence intervals in all cases and that the Wald interval is not always a subset of the AR interval. For example, in Column (4) with the latitude control, the Wald confidence interval is [0.50, 1.08], whereas the AR interval is [0.53, 1.67]. Even though, the effective F-statistic is 19.7 in this case, which rejects instrument weakness at 5% and 10% significance levels under 20% and 10% tolerance bias, respectively, the cluster-robust AR confidence interval is 93% larger than the

[5] The original Acemoglu et al. (2001) study used the Wald statistic for inference, without any adjustment for heteroskedasticity.

Wald confidence interval.

The AR bootstrapped confidence intervals are of the same or smaller magnitude than the AR asymptotic confidence intervals when the effective F-tests are larger. For example, Columns (1) to (3) of the specification without African countries have effective F-tests equal to or above 23, which is approximately the 5% critical value of this test with a tolerance bias of 10%.

On the other hand, with an effective F-test below 12.4, which is the 10% critical value under 20% bias tolerance, the AR bootstrapped confidence intervals in general have larger lengths than the asymptotic ones do, as illustrated by the specifications with continent dummies and latitudes, those with the percentage of European descent in 1975, and those with malaria. In most of these specifications, the asymptotic AR confidence intervals are a segment located on the positive side of the real line, whereas the bootstrapped confidence intervals stretch to negative values as well. These latter results indicate that the effect of institutions on growth is not statistically significant. We even observe cases with disjoint AR bootstrapped confidence intervals, as in columns (3) and (4) for the specification with malaria.

Finally, we observe that the Wald bootstrapped confidence intervals are larger than their asymptotic counterparts. These differences can also be substantial when the instrument is undoubtedly strong, as suggested by the effective F-test. This is illustrated by the specifications with no covariates and without African countries in Columns (1) and (4), which have effective F-tests above 23. The Wald bootstrapped confidence intervals are at least 27% larger than the asymptotic ones. When the instruments are weak, and, consequently, inference with Wald tests is unreliable, the differences are even more pronounced.

To further investigate the statistical significance of institutions on growth we report the *p-values* (multiplied by 100) for testing $H_0 : \theta = 0$ across the specifications on Table 2. The lightest shade of gray indicates if the test is not rejected at the 1% significance level, the medium gray at the 5% level, the medium-dark at the 10% level, and the darkest at the 20% level.

[Table 2 about here.]

Focusing firstly on the asymptotic AR test, the majority of specifications indicate significance of the institutional effect on growth at the 1% significance level. When we include Albouy's "campaign" dummy the same effect becomes statistically insignificant at the same level. Using the extended correction to Albouy's "campaign" dummy proposed by Acemoglu et al. restores the significance of the institution effect. There is an increasing pattern from the AR to the bootstrapped AR *p-values*, which shows a less significant effect of institutions on growth. For example, the specifications with the extended correction to the Albouy campaign dummy, the effect of institutions on growth is significant using the asymptotic AR test at the 1% significance level for most cases, but insignificant at the same level using the AR bootstrap test.

We can draw one important takeaway from this application: tests based on the asymptotic critical values can misguide inference. As we would expect, the weaker the instrument, the larger the confidence intervals, and the larger the differences between the asymptotic and bootstrapped confidence intervals. With effective F-tests below 12.4, however, we can still estimate bounded confidence intervals and obtain useful information about the structural parameter. The misguidance provided by asymptotic tests can also occur even when the instruments are strong according to the effective F-test.

3.2 Application 2: Civil War

The second application comes from Miguel et al. (2004). This example is interesting because the AR statistic uses a null hypothesis with joint restrictions on two endogenous variables. The resulting confidence sets are two-dimensional areas that need not be convex. A challenge with this application is that the literature has not yet developed a weak instrument test similar to the one from (Olea and Pflueger, 2013) for the case of multiple endogenous variables. However, we can still rely on the AR test to conduct inference of the structural parameter whether or not instruments are weak.

Miguel et al. (2004) investigate the relationship between economic conditions and civil war in sub-Saharan Africa. In particular, they study how the deterioration of the economic environment affects the probability of civil conflict. Endogeneity bias may arise from the simultaneity of government institution quality, economic performance, and civil war. The authors use variation in rainfall as an instrument for economic performance, which is salient because of the reliance on subsistence agriculture in the sample countries. The data consist of an unbalanced panel of 41 African countries from 1981 to 1999, with 743 total observations, averaging 18.6 observations per country. The primary models have a binary measure of civil war as an outcome, current and lagged economic performance as endogenous variables, and current and lagged rainfall growth as instruments.[6] The original study does not report tests that are robust to the presence of weak instruments.

Figure 1 shows asymptotic and bootstrapped Wald and AR confidence sets for three specifications in Miguel et al. (2004). The left column shows Wald confidence sets, and the right column shows the corresponding AR sets. The gray areas are the 95% asymptotic confidence sets, and the

[6] All of the models that we replicate include country fixed effects and country-specific time trends.

black lines show the bootstrapped 95% confidence sets. In the figures, θ_1 is the current economic growth rate, and θ_2 is the lagged economic growth rate. The parameter estimates are the center points of the asymptotic Wald sets. The specifications only differ with respect to the dependent variables.

Figure 1 also reports the Kleibergen and Paap (KP) rank test (Kleibergen and Paap, 2006). The KP test tests the hypothesis that Π_z in Equation (1), which is a matrix in the case of multiple endogenous variables, has reduced rank. If the KP test is not rejected, then θ is underindentified. On the other hand, if we reject the KP test, we cannot conclude that the instruments are strong.[7]

The first row of Figure 1 is a replication of Table 6, Column 1, in which the dependent variable is the PRIO/Uppsala indicator for the onset of civil conflict. The bootstrapped Wald confidence set is somewhat larger than the asymptotic set. Although the KP rank is rejected at the 0.01% significance level, the instruments are relatively weak, which is reflected in a much larger asymptotic AR confidence set. Once we account for the cluster structure by bootstrapping, the AR confidence set blows up. Accounting for the combination of weak instruments and a moderately small number of clusters makes inference difficult.

The second row of Figure 1 is a replication of Table C3, Column 4, in which the dependent variable is the Doyle and Sambanis indicator for periods with major civil conflicts (more than 1,000 deaths). In this specification, the KP rank statistic is larger. Bootstrapping the Wald statistic leads to a similar expansion of the confidence set. For the AR test, bootstrapping causes expansion, but here the expansion is all in one direction in the parameter space—toward a negative estimate of θ_1 and a positive estimate of θ_2, which has no meaningful economic interpretation and is inconsistent with the original paper. The starkest take-away from the confidence sets is the expansion to implausibly large parameter values.

The third row of Figure 1 is a replication of Table C3, Column 5, which has the same dependent variable concepts but uses the Fearon and Laitin measure of civil conflict. Here we see interesting patterns in how inference is affected by the use of joint null hypotheses. The bootstrapped Wald confidence set expands only for θ_1, and mainly in a direction that no longer supports the paper's conclusions. For the AR set, the uncertainty about θ_2 depends greatly on the value of θ_1.

[Figure 1 about here.]

In Section S.1 of the Supplement, we report results for the first-stage F-test, the (projected)

[7] Andrews et al.'s (2018) survey shows that practitioners compute the KP test and compare it to the critical values of the Stock and Yogo (2005) weak instrument test, because the KP test reduces to the (robust) first-stage F-test in the case of one endogenous variable.

confidence intervals, and the *p-values* of the joint statistical significant of $\theta = (\theta_1, \theta_2)$. Overall they indicate the statistical insignificance of the effect of economic conditions on civil war, which is contrary to the findings of the original paper.

4 Testing structural parameters with clustered errors and weak instruments

We now present the general case of the Wald and AR tests derived from models with more than one endogenous variable and several included and excluded instruments. The bootstrap versions of these tests are in the following section.[8]

With multiple endogenous explanatory variables and exogenous controls, the general representation of simple cluster model (1) becomes

$$\begin{cases} \mathbf{y}_{1,g} = & \mathbf{y}_{2,g}\theta + \mathbf{x}_g\gamma + \mathbf{u}_g \\ \mathbf{y}_{2,g} = & \mathbf{w}_g\Pi_w + \mathbf{v}_g \end{cases} \quad \text{for } g = 1, \ldots, G, \tag{3}$$

where $\mathbf{y}_{1,g}$ and $\mathbf{u}_g$ are $n_g \times 1$ vectors, $\mathbf{y}_{2,g}$ and $\mathbf{v}_g$ are $n_g \times p$ matrices, $\mathbf{w}_g = [\mathbf{z}_g \colon \mathbf{x}_g]$ is a $n_g \times k_w$ matrix of instruments, $\mathbf{z}_g$ and $\mathbf{x}_g$ are $n_g \times k_z$ and $n_g \times k_x$ matrices of excluded and included instruments with $k_w = k_z + k_x$, and $\Pi_w = [\Pi_z' \, \Pi_x']'$ is a $k_w \times p$ matrix of first-stage, reduced-form parameters. The errors $(\mathbf{u}_g, \mathbf{v}_g)$ are independent across clusters with variance $\mathrm{E}\left[(\mathbf{u}_g, \mathrm{vec}\,(\mathbf{v}_g))\,(\mathbf{u}_g, \mathrm{vec}\,(\mathbf{v}_g))'\right] = \mathbf{\Sigma}_g$. The equations in (3) have the following matrix representation

$$\begin{cases} \mathbf{y}_1 = & \mathbf{y}_2\theta + \mathbf{X}\gamma + \mathbf{u} \\ \mathbf{y}_2 = & \mathbf{W}\Pi_w + \mathbf{V}, \end{cases} \tag{4}$$

where $\mathbf{y}_1$ is an $n \times 1$ vector, $\mathbf{y}_2$ is an $n \times p$ matrix of endogenous explanatory variables, $\mathbf{W} = [\mathbf{Z} : \mathbf{X}]$ is an $n \times k_w$ matrix of instruments, and $\mathbf{Z}$ and $\mathbf{X}$ are $n \times k_z$ and $n \times k_x$ matrices of excluded and included instruments, respectively. The Wald test is defined as

$$W(\theta_0) = \left(\hat{\theta}_{\mathrm{IV}} - \theta_0\right)' \left(\widehat{\mathrm{Var}}(\hat{\theta}_{\mathrm{IV}})\right)^{-1} \left(\hat{\theta}_{\mathrm{IV}} - \theta_0\right), \tag{5}$$

where $\hat{\theta}_{\mathrm{IV}} = \left(\mathbf{y}_2'\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}}\mathbf{y}_2\right)^{-1} \mathbf{y}_2'\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}}\mathbf{y}_1$ is the IV estimator and $\widehat{\mathrm{Var}}(\hat{\theta}_{\mathrm{IV}})$, the cluster-robust estima-

[8] The definitions of the cluster-robust versions of the KLM and CLR tests together with the their bootstrap counterparts can be found in the Supplement.

tor of $\mathrm{Var}(\hat{\theta}_{\mathrm{IV}})$, is

$$\widehat{\mathrm{Var}}\left(\hat{\theta}_{\mathrm{IV}}\right)=\left(\mathbf{y}_2'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{y}_2\right)^{-1}\left[\sum_{g=1}^{G}\left(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{y}_2\right)_g'\,\widehat{\boldsymbol{\Sigma}}_g(\hat{\theta}_{\mathrm{IV}})\left(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{y}_2\right)_g\right]\left(\mathbf{y}_2'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{y}_2\right)^{-1}, \tag{6}$$

where $(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{y}_2)_g$ is the $n_g \times p$ submatrix $\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{y}_2$ associated with the g^{th} cluster.

Tests robust to the presence of weak instruments

The AR, KLM, and CLR tests were originally developed under the assumption that the distribution of the errors is iid, but they have been adapted to allow for arbitrary heteroskedasticity or cluster dependence of the residuals (Chernozhukov and Hansen, 2008; Finlay and Magnusson, 2009). We start by redefining the equations in (4) as

$$\begin{cases}\mathbf{Y}(\theta_0)= & \mathbf{W}\delta_w(\theta_0)+\mathbf{e}(\theta_0)\\ \mathbf{y}_2= & \mathbf{W}\Pi_w+\mathbf{V},\end{cases} \tag{7}$$

where $\mathbf{Y}(\theta_0)=\mathbf{y}_1-\mathbf{y}_2\theta_0$, $\mathbf{e}(\theta_0)=\mathbf{u}+\mathbf{V}\mathrm{d}(\theta_0)$, $\delta_w(\theta_0)=\left[\delta_z(\theta_0)',\delta_x(\theta_0)'\right]'=\Pi_w\mathrm{d}(\theta_0)+\mathrm{H}\gamma$, $\Pi_w=[\Pi_z',\Pi_x']'$, $\mathrm{d}(\theta_0)=(\theta-\theta_0)$, and $\mathrm{H}=[0,\mathrm{I}_{k_x}]'$. The first equations in (7) can be further rewritten as

$$\hat{\delta}_w(\theta_0)=\underbrace{\Pi_w\mathrm{d}(\theta_0)+\mathrm{H}\gamma}_{\delta_w(\theta_0)}+(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{e}(\theta_0) \tag{8}$$

where $\hat{\delta}_w(\theta_0)=[\hat{\delta}_z(\theta_0)',\hat{\delta}_x(\theta_0)']'=(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}(\theta_0)$ is the OLS estimator of the reduced-form parameter $\delta_w(\theta_0)$. The $k_w \times k_w$ "sandwich" matrix that corresponds to the cluster-robust estimator of the variance of $\hat{\delta}_w(\theta_0)$ is

$$\widehat{\Omega}(\theta_0)=\left(\mathbf{W}'\mathbf{W}\right)^{-1}\widehat{\Xi}_{ee}(\theta_0)\left(\mathbf{W}'\mathbf{W}\right)^{-1}, \tag{9}$$

where $\widehat{\Xi}_{ee}(\theta_0)$ is the estimator of the $k_w \times k_w$ variance matrix of $\mathbf{W}'\mathbf{e}(\theta_0)$.[9]

Definition 1 (AR test with clustered residuals)**.** The AR statistic for testing the null hypothesis $H_0:\mathrm{d}(\theta_0)=0$ is:

$$\Lambda_{\mathrm{AR}}(\theta_0)\equiv\hat{\delta}_z(\theta_0)'\left[\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)\right]^{-1}\hat{\delta}_z(\theta_0)\xrightarrow{d}\chi^2(k_z),$$

where $\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)$ is the submatrix of $\widehat{\Omega}(\theta_0)$ associated with the variance and covariance estimator of $\hat{\delta}_z(\theta_0)$. The symbol $\xrightarrow{d}$ represents convergence in distribution as $G\to+\infty$, and $\chi^2(s)$ is the chi-squared distribution with s degrees of freedom.

[9] Details about the computation of $\widehat{\Xi}(\theta_0)$ are in Section S.4 of the Supplement.

The AR test has the correct asymptotic size even when the structural parameter θ is not identified. In that case, the AR test will not have power, indicating the presence of weak instruments. On the other hand, the AR test is consistent when $\Pi_z \gg 0$.[10] The number of excluded instruments, k_z, corresponds with the degrees of freedom of the AR test distribution, which can be larger than p, the number of tested parameters. The larger the difference $k_z - p$, the less powerful the AR test. The AR test also has a Lagrange-multiplier interpretation—see Section S.3 in the Supplement.

5 Bootstrap methods for the cluster-sample IV model

In many microeconometric applications, data have intra-cluster dependence in which the number of clusters is small, and the asymptotic results are consequently poor approximations of the true distributions of the test statistics. For example, many studies in labor economics use research designs that rely on policy changes at the state level, in which the number of clusters is at most 51 in the USA and eight in Australia. Our simulations show that asymptotic tests that use cluster-robust variance estimators may overreject with as many as 80 clusters. Therefore, bootstrapping them accordingly can improve their size performance when the number of clusters is small.

We next present two classes of bootstrap methods for the AR test in a linear IV cluster model represented by System (3): the estimating equations and the residual bootstraps.

5.1 Estimating equations (score) bootstrap

We begin the exposition by rewriting Equation (8) as

$$\hat{\delta}_w(\theta_0) = \delta_w(\theta_0) + \left(\mathbf{W}'\mathbf{W}\right)^{-1} \sum_{g=1}^{G} \underbrace{\mathbf{w}_g' \mathbf{e}_g(\theta_0)}_{\mathbf{h}_g(\theta_0)}.$$

A simple idea about bootstrapping the distributions of $\hat{\delta}_w(\theta_0)$ is based on perturbing the empirical distribution of the scores $\{\mathbf{h}_g(\theta_0)\}_{g=1}^{G}$, but keeping the Hessian $(\mathbf{W}'\mathbf{W})^{-1}$ fixed. Hu and Zidek (1995) denote this type of bootstrap the estimating equations (EE) bootstrap.[11]

Under $H_0 : \mathrm{d}(\theta_0) = 0$, a candidate bootstrap estimator for $\delta_w(\theta_0)$ is

$$\tilde{\delta}_w^*(\theta_0) = \tilde{\delta}_w(\theta_0) + \left(\mathbf{W}'\mathbf{W}\right)^{-1} \sum_{g=1}^{G} \tilde{\mathbf{h}}_g^*(\theta_0), \tag{10}$$

[10] A test is consistent if it rejects $H_0 : \mathrm{d}(\theta_0) = 0$ when $H_1 : \mathrm{d}(\theta_0) \neq 0$ is true and the sample size increases.

[11] See also Hu and Kalbfleisch (2000) and Kline and Santos (2012).

where $\tilde{\delta}_w(\theta_0) = (0, \tilde{\delta}_x(\theta_0))$ and $\tilde{\delta}_x(\theta_0) = \hat{\delta}_x(\theta_0) - \widehat{\Omega}_{\delta_x\delta_z}(\theta_0)\left[\widehat{\Omega}_{\delta_z\delta_z}\right]^{-1}\hat{\delta}_z(\theta_0)$. $\tilde{\delta}_w(\theta_0)$ is a minimum distance estimator of $\delta_w(\theta_0)$—see Section S.3 in the Supplement. The sequence of bootstrap scores $\left\{\tilde{\mathbf{h}}_g^*(\theta_0)\right\}_{g=1}^G$ is sampled with replacement from the recentered scores $\{\tilde{\mathbf{h}}_g^r(\theta_0)\}_{g=1}^G$, defined as

$$\tilde{\mathbf{h}}_g^r(\theta_0) = \tilde{\mathbf{h}}_g(\theta_0) - \frac{n_g}{n}\sum_{j=1}^G \tilde{\mathbf{h}}_j(\theta_0),$$

where $\tilde{\mathbf{h}}_g(\theta_0) = \mathbf{w}_g'\tilde{\mathbf{e}}_g(\theta_0)$ and $\tilde{\mathbf{e}}_g(\theta_0) = \mathbf{Y}_g(\theta_0) - \mathbf{w}_g\tilde{\delta}_w(\theta_0)$.[12]

The estimator of the variance of $\tilde{\delta}_w^*(\theta_0)$, denoted by $\widetilde{\Omega}_{\delta_w\delta_w}^*(\theta_0)$, is a function of $\left\{\tilde{\mathbf{h}}_g^*(\theta_0)\right\}_{g=1}^G$ and does not depend on $\tilde{\delta}_w^*(\theta_0)$ itself. This property implies a computational gain of the EE bootstrap over the residual-type bootstraps discussed below.

The general algorithm for computing the bootstrap AR test is:

1. Compute $\Lambda_{\text{AR}}(\theta_0)$

2. For $b = 1, \ldots, B$ bootstrap simulations:

   (a) Sample $\{\omega_g\}_{g=1}^G$, a sequence of bootstrap weights defined using Definition 2 below, and set the bootstrap score realizations as:

   $$\left\{\tilde{\mathbf{h}}_1^*(\theta_0), \ldots, \tilde{\mathbf{h}}_G^*(\theta_0)\right\} = \left\{\omega_1\tilde{\mathbf{h}}_1^r(\theta_0), \ldots, \omega_G\tilde{\mathbf{h}}_G^r(\theta_0)\right\}.$$

   (b) Compute $\tilde{\delta}_w^*(\theta_0)$ given by Equation (10) and its associated variance $\widetilde{\Omega}_{\delta_w\delta_w}^*(\theta_0)$.

   (c) The b^{th} bootstrap test is

   $$\tilde{\Lambda}_{\text{AR},b}^*(\theta_0) = \tilde{\delta}_z^*(\theta_0)'\left[\widetilde{\Omega}_{\delta_z\delta_z}^*(\theta_0)\right]^{-1}\tilde{\delta}_z^*(\theta_0), \tag{11}$$

   where $\widetilde{\Omega}_{\delta_z\delta_z}^*(\theta_0)$ is the block variance of $\widetilde{\Omega}_{\delta_w\delta_w}^*(\theta_0)$ associated with $\tilde{\delta}_z^*(\theta_0)$.

3. The bootstrap p-value for the AR test is:

   $$\tilde{p}^*\text{-value} = \frac{1}{B}\sum_{b=1}^B I\left(\tilde{\Lambda}_{\text{AR},b}^*(\theta_0) > \Lambda_{\text{AR}}(\theta_0)\right),$$

   where $I(\cdot)$ is the indicator function. Reject the null hypothesis if the $\tilde{p}^*$-value is smaller than the desired significance level of the test.

Next, we discuss two types of estimating equation bootstraps.

[12] If the number of observations per cluster is the same, then $\frac{n_g}{n} = \frac{1}{G}$.

Definition 2 (Estimating equation (EE) bootstrap). Let $\{\omega_g\}_{g=1}^{G}$ be a sequence of bootstrap weights. The AR EE bootstrap test is computed from the bootstrap score sequence $\left\{\tilde{\mathbf{h}}_g^* (\theta_0)\right\}_{g=1}^{G} = \{\omega_g \tilde{\mathbf{h}}_g^r (\theta_0)\}_{g=1}^{G}$. We consider two bootstrap weights:

1. $\{\omega_g\}_{g=1}^{G}$ are sampled from a multinomial distribution, so that

$$\Pr\left(\tilde{\mathbf{h}}_g^* (\theta_0) = \tilde{\mathbf{h}}_j (\theta_0)\right) = \frac{1}{G}, \quad j = 1, \ldots, G.$$

2. $\{\omega_g\}_{g=1}^{G}$ is an iid sequence sampled from a distribution satisfying $E[\omega_g] = 0$ and $\mathrm{Var}(\omega_g) = 1$. We discuss the specific distributions for $\{\omega_g\}_{g=1}^{G}$ below.

The EE bootstrap with multinomial weights is closely related to bootstrap algorithm 1 of Kleibergen (2011) for GMM models. The second bootstrap is similar to the wild score bootstrap method proposed by Kline and Santos (2012). These authors assume, however, that the tested parameter is identified and consistently estimated, which allows them to use two-step GMM estimates on the empirical score. Clearly, the GMM estimator is inconsistent when the instruments are weak.

Remark 5.1. Sampling the score from $\left\{\omega_g \tilde{\mathbf{h}}_g^r (\theta_0)\right\}_{g=1}^{G}$ corresponds to sampling the residuals from $\{\bar{\omega}_g \tilde{\mathbf{e}}_g (\theta_0)\}_{g=1}^{G}$, where $\bar{\omega}_g = \omega_g - \bar{\omega}$ and $\bar{\omega} = \left(\sum_{g=1}^{G} \frac{\omega_g n_g}{n}\right)$. We can interpret $\{\bar{\omega}_g\}_{g=1}^{G}$ as the sequence of adjusted bootstrap weights.

Remark 5.2. The estimator $\acute{\delta}_w (\theta_0) = (0, \acute{\delta}_x (\theta_0))$, where $\acute{\delta}_x (\theta_0) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} (\theta_0)$, could replace $\tilde{\delta}_w (\theta_0)$ in Equation (10). The bootstrap scores would be generated as before, and recentering is unnecessary if a constant is included in $\mathbf{x}_g$. If the OLS estimate of $\hat{\delta}_w (\theta_0)$ replaces $\tilde{\delta}_w (\theta_0)$ in the bootstrap, then $\tilde{\delta}_z^* (\theta_0)$ in Equation (11) should be substituted with $\hat{\delta}_z^* (\theta_0) - \hat{\delta}_z (\theta_0)$ because $\hat{\delta}_z (\theta_0)$ is the mean of the distribution of the bootstrap estimator $\hat{\delta}_z^* (\theta_0)$.

5.2 Residual bootstraps

We can also resample the estimated residuals to generate bootstrap samples for obtaining bootstrap test statistics. We consider two types of residual bootstraps. The first version estimates the bootstrap DGP by resampling the residuals of the second stage regression in (7). In the second version, we simultaneously resample residuals from both the first and second-stage regressions for estimating the bootstrap DGP. We refer to the first and second bootstrap methods as single-equation (SE) and multi-equation (ME) residual bootstraps. The bootstrap test statistics are computed in the same way as the asymptotic ones, replacing the original sample with the bootstrapped one.

The AR-SE, and AR-ME bootstrap tests are equal because the AR test does not depend on the first-stage regression residuals. Since we focus on the AR test, we present only the single-equation residual bootstrap. We describe in the Supplement multi-equation residual bootstrap methods for the KLM, CLR, and Wald tests, and a variant of Davidson and MacKinnon's (2010) bootstrap residual procedure for the AR and KLM tests adapted to the cluster case.

Single-equation residual bootstrap

Let $\{\acute{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$ be a sequence of residuals, where $\acute{\mathbf{e}}_g(\theta_0) = \mathbf{Y}_g(\theta_0) - \mathbf{w}_g\acute{\delta}_w(\theta_0)$ is the $n_g \times 1$ vector associated to the g^{th} cluster and $\acute{\delta}_w(\theta_0)$ is defined in Remark 5.2. Then, the bootstrap realization of $\mathbf{Y}(\theta_0)$ is given by $\acute{\mathbf{Y}}^*(\theta_0) = \mathbf{W}\acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}^*(\theta_0)$, where $\acute{\mathbf{e}}^*(\theta_0) = (\acute{\mathbf{e}}_1^*(\theta_0)', \ldots, \acute{\mathbf{e}}_G^*(\theta_0)')'$, $\left\{\acute{\mathbf{e}}_g^*(\theta_0)\right\}_{g=1}^G = \{\omega_g\acute{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$, and the bootstrap estimates of $\delta_w(\theta_0)$ are

$$\acute{\delta}_w^*(\theta_0) = \left[\begin{array}{cc} \acute{\delta}_z^*(\theta_0)' & \acute{\delta}_x^*(\theta_0)' \end{array}\right]' = \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\acute{\mathbf{Y}}^*(\theta_0).$$

The steps for implementing the AR *residual* bootstrap test are similar to those for the AR EE bootstrap test. In Step 2a, we use a sequence of $\left\{\acute{\mathbf{e}}_g^*(\theta_0)\right\}_{g=1}^G = \{\omega_g\acute{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$. In Steps 2a and 2b, we compute $\acute{\delta}_w^*(\theta_0)$ and $\acute{\Omega}_{\delta_z^*\delta_z^*}^*(\theta_0)$, where $\acute{\Omega}_{\delta_z^*\delta_z^*}^*(\theta_0)$ is an estimator of the variance matrix $\acute{\delta}_z^*(\theta_0)$.[13] This result differs from the EE bootstrap in that $\acute{\Omega}_{\delta_z^*\delta_z^*}^*(\theta_0)$ is a function of $\acute{\delta}_z^*(\theta_0)$. The b^{th} bootstrap $\acute{\Lambda}_{\text{AR},b}^*(\theta_0)$ is obtained by replacing $\tilde{\delta}_z^*(\theta_0)$ and $\widetilde{\Omega}_{\delta_z\delta_z}^*(\theta_0)$ with $\acute{\delta}_w^*(\theta_0)$, $\acute{\Omega}_{\delta_z^*\delta_z^*}^*(\theta_0)$ in the respective formulas of Step 2c.

Definition 3 (Single-equation residual (SE) bootstrap)**.** Let $\{\omega_g\}_{g=1}^G$ be a sequence of bootstrap weights satisfying $E[\omega_g] = 0$ and $\text{Var}(\omega_g) = 1$. The bootstrap DGP is:

1. Inefficient SE (SE-in):
   $\left\{\acute{\mathbf{Y}}_g^*(\theta_0)\right\}_{g=1}^G = \left\{\mathbf{w}_g\acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}_g^*(\theta_0)\right\}_{g=1}^G$, where $\left\{\acute{\mathbf{e}}_g^*(\theta_0)\right\}_{g=1}^G = \{\omega_g\acute{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$, and
2. Efficient SE (SE-eff):
   $\left\{\tilde{\mathbf{Y}}_g^*(\theta_0)\right\}_{g=1}^G = \left\{\mathbf{w}_g\tilde{\delta}_w(\theta_0) + \tilde{\mathbf{e}}_g^*(\theta_0)\right\}_{g=1}^G$, where $\left\{\tilde{\mathbf{e}}_g^*(\theta_0)\right\}_{g=1}^G = \{\omega_g\tilde{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$.

Remark 5.3. If a constant is not included in $\mathbf{x}_g$, then the fitted residuals $\{\acute{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$ should be recentered.

Remark 5.4. As in Remark 5.2, we could use $\left\{\hat{\mathbf{e}}_g^*(\theta_0)\right\}_{g=1}^G = \{\omega_g\hat{\mathbf{e}}_g(\theta_0)\}_{g=1}^G$, where $\hat{\mathbf{e}}_g = \mathbf{Y}_g(\theta_0) - \mathbf{w}_g\hat{\delta}_w(\theta_0)$, to generate bootstrap realizations of $\mathbf{Y}(\theta_0)$. Then, $\hat{\delta}_z^*(\theta_0) - \hat{\delta}_z(\theta_0)$ should be in place

[13] See Section S.4 in the Supplement for the definition of the variance estimator.

of $\hat{\delta}_z^*(\theta_0)$ when computing the bootstrap realizations of the AR test, where $\hat{\delta}_z^*(\theta_0)$ is the bootstrap estimator of $\hat{\delta}_z(\theta_0)$. In this case, the only difference between the EE and SE bootstraps is the bootstrap estimator of the variance. For the EE bootstrap, we use $\left\{\hat{\mathbf{e}}_g^*(\theta_0)\right\}_{g=1}^G$, whereas for the SE bootstrap, we use $\{\hat{\mathbf{e}}_{\mathrm{b},g}^*(\theta_0)\}_{g=1}^G$, where $\hat{\mathbf{e}}_{\mathrm{b},g}^*(\theta_0) = \hat{\mathbf{Y}}_g^*(\theta_0) - \mathbf{w}_g\hat{\delta}_w^*(\theta_0)$.

5.3 Resampling weights

Apart from the EE bootstrap with multinomial weights, the remaining weights used for the proposed bootstraps satisfy $E[\omega_g] = 0$ and $E[\omega_g^2] = 1$. This condition ensures that the distributions of the resampled scores or residuals have the same first and second moments of their underlying empirical distributions. Matching higher moments of the bootstrap and empirical distributions yields the asymptotic refinement. Many weights satisfy this property for the wild bootstrap. Liu (1988) proposes weights defined as $\omega_g = \zeta_g - E(\zeta_g)$, where ζ_g is a gamma random variable with shape parameter 4 and scale parameter $\frac{1}{2}$. The gamma (Γ) weights also satisfy $E[\omega_g^3] = 1$ and, therefore, match the first three moments. Davidson and MacKinnon (2010) suggest sampling the weights from the Rademacher distribution, which is defined as

$$\omega_g = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}.$$

The Rademacher (R) weights match the first four moments if the underlying distribution is symmetric.

There are other bootstrap weights based on continuous and discrete distributions, such as Liu's weights ($\omega_g = w_g z_g - E(w_g)E(z_g)$, where w_g and z_g are independent normal random variables with mean $\frac{1}{2}(\sqrt{17/6}+\sqrt{1/6})$ and variance $\frac{1}{2}$), and Mammen's weights ($\omega_g = (1-\sqrt{5})/2$ with probability $\frac{1+\sqrt{5}}{2\sqrt{5}}$ and $\omega_g = (1+\sqrt{5})/2$ with probability $\frac{\sqrt{5}-1}{2\sqrt{5}}$). Our Monte Carlo experiments show that Liu's and Mammen's weights are outperformed by the gamma and Rademacher weights, so we only report the results using gamma and Rademacher weights.[14]

6 Monte Carlo simulations

We now evaluate the small sample properties of the asymptotic and proposed cluster bootstrap tests through Monte Carlo simulations.

[14] The results are available upon request.

6.1 Simulation Design

The baseline model has the same structure as System (3), which is repeated for convenience:

$$\begin{cases} \mathbf{y}_{1,g} = & \mathbf{y}_{2,g}\theta + \mathbf{x}_g\gamma + \mathbf{u}_g \\ \mathbf{y}_{2,g} = & \mathbf{z}_g\Pi_z + \mathbf{x}_g\Pi_x + \mathbf{v}_g \end{cases} \quad \text{for } g = 1, \ldots, G.$$

We assume that θ is scalar; $\mathbf{x}_g = \boldsymbol{\iota}_g$, where $\boldsymbol{\iota}_g$ is a $n_g \times 1$ vector of ones; and $\gamma = \Pi_x = 1$. The instrument $\mathbf{z}_g$ is set as $\mathbf{z}_g = \iota_{n_g}\mathbf{d}'_g + \boldsymbol{\vartheta}_g$, where $\mathbf{d}_g$ is a $k_z \times 1$ vector and $\boldsymbol{\vartheta}_g$ is an $n_g \times k_z$ matrix and where $\mathbf{d}_g$ and $\boldsymbol{\vartheta}_g$ are sampled from independent multivariate distributions and adjusted such that $\sum_{g=1}^{G} n_g \left(\mathbf{d}_g - \overline{\mathbf{d}}\right)' \left(\mathbf{d}_g - \overline{\mathbf{d}}\right) = (1-\lambda)\, n\mathrm{I}_{k_z}$, where $\overline{\mathbf{d}} = \frac{1}{G}\sum_{g=1}^{G} \mathbf{d}_g$ and $\sum_{g=1}^{G} \boldsymbol{\vartheta}'_{n_g}\boldsymbol{\vartheta}_{n_g} = \lambda\, n\mathrm{I}_{k_z}$, with $\boldsymbol{\iota}'_g\boldsymbol{\vartheta}_g = 0$. These adjustments allow us to have $n^{-1}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z} = \mathrm{I}_{k_z}$, where n is the total number of observations. The scalar λ satisfies $0 \leq \lambda \leq 1$. If $\lambda = 0$, then the instruments are the same within groups. We keep the instruments $\mathbf{W} = [\mathbf{Z} : \mathbf{X}]$ fixed in all simulations.

Carter et al. (2016) and MacKinnon and Webb (2014) show that variation in the number of observations across clusters does affect hypothesis testing in the linear regression with only exogenous controls. Therefore, for a fixed sample size n, we follow MacKinnon and Webb and set the number of observations in cluster g according to

$$n_g = \mathrm{nint}\left(n\frac{\exp\left(\eta\frac{g}{G}\right)}{\sum_{g=1}^{G}\exp\left(\eta\frac{g}{G}\right)}\right), \text{ for } g = 1, \ldots, G-1, \text{and } \; n_G = n - \sum_{g=1}^{G-1} n_g,$$

where $\mathrm{nint}(\cdot)$ is the nearest integer function. When $\eta = 0$, each cluster has the same number of observations. As η increases, the differences across the n_gs are larger.

We generate the model errors as

$$\begin{cases} \mathbf{u}_g & = \mathbf{u}^c_g + \mathbf{u}^i_g \\ \mathbf{v}_g & = \rho\mathbf{u}^c_g + \varrho\mathbf{u}^i_g + \left(1-\rho^2\right)^{\frac{1}{2}}\mathbf{v}^c_g + \left(1-\varrho^2\right)^{\frac{1}{2}}\mathbf{v}^i_g \end{cases} \quad \text{for} \quad g = 1, \ldots, G,$$

where $\left(\mathbf{u}^c_g, \mathbf{v}^c_g\right) = \sqrt{\phi}\left(\varepsilon_{1,g}, \varepsilon_{2,g}\right) \otimes \boldsymbol{\iota}_g$, $\left(\mathbf{u}^i_g, \mathbf{v}^i_g\right) = \sqrt{1-\phi}(\boldsymbol{\psi}_{1,g} \odot f\left(\mathbf{z}_{1,g}, \kappa\right), \boldsymbol{\psi}_{2,g})$, $\boldsymbol{\varepsilon}_g = (\varepsilon_{1,g}, \varepsilon_{2,g}) \sim iid\left(0, \mathrm{I}_2\right)$, $\boldsymbol{\psi}_g = (\boldsymbol{\psi}'_{1,g}, \boldsymbol{\psi}'_{2,g})' \sim iid(0,\mathrm{I}_{2n_g})$, and $\boldsymbol{\varepsilon}_g$ and $\boldsymbol{\psi}_g$ are independent. The operator $\odot$ is the Hadamard product, and $f\left(\mathbf{z}_{1,g}, \kappa\right)$ is the skedastic function, which is defined as

$$f\left(\mathbf{z}_{1,g}, \kappa\right) = h\left(\kappa\right)\left(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g}\right)^{\kappa}, \tag{12}$$

where the vector $\mathbf{z}_{1,g}$ corresponds to the first column of the matrix $\mathbf{z}_g$. The function $h\left(\kappa\right)$ is a scaling factor that ensures that the average variance of $\mathbf{u}^i$ is equal to $1-\phi$. Therefore, changing

κ captures changes in the heterogeneity present in the errors without affecting the variance of these errors on average. When $\kappa = 0$, the residuals are akin to the cluster random-effects model. The scalar ϕ, which satisfies $0 \leq \phi \leq 1$, has two roles. First, it captures the correlation of disturbances within the same cluster. Second, it controls the share of the variance of $\mathbf{u}_g$ due to idiosyncratic ($\mathbf{u}_g^i$) and cluster ($\mathbf{u}_g^c$) components. The scalars ρ and ϱ capture the intra-cluster and the idiosyncratic correlations, respectively.

The identification of the structural parameter θ is related to the rank of Π_z. We set Π_z based on the noncentrality parameter of the cluster-robust first-stage F-statistic for testing $H_0 : \Pi_z = 0$, which is

$$\mu_{k_z} = n \frac{\Pi_z' \left[\mathrm{Var}_\infty\left(\hat{\Pi}_z\right)\right]^{-1} \Pi_z}{k_z},$$

where $\mathrm{Var}_\infty(\hat{\Pi}_z) = \lim_{n \to +\infty} \frac{1}{n} \mathrm{E}[\mathbf{Z}' \mathrm{M}_{\mathbf{X}} \mathbf{V} \mathbf{V}' \mathrm{M}_{\mathbf{X}} \mathbf{Z}]$. If residuals are independent across individuals and homoskedastic, then μ_{k_z} becomes the "concentration parameter" divided by the number of excluded instruments. The inverse of the noncentrality parameter μ_{k_z} also appears in the formula of the bias of the IV estimator derived from a Nagar type of expansion (see Equation (15) below and Sections S.6 and S.7 in the Supplement). We assume that only the first instrument is relevant, so that $\Pi_z = (c_z, 0, ..., 0)'$. Therefore, the noncentrality parameter simplifies to

$$\mu_{k_z} = n \frac{c_z^2}{k_z} \left[\mathrm{Var}_\infty\left(\hat{\Pi}_z\right)\right]_{11}^{-1}, \tag{13}$$

where $\left[\mathrm{Var}_\infty(\hat{\Pi}_z)\right]_{11}^{-1}$ indicates the first diagonal entry of $\left[\mathrm{Var}_\infty(\hat{\Pi}_z)\right]^{-1}$, and we fix the value of c_z as

$$c_z = \sqrt{\frac{k_z}{\left[\mathrm{Var}_\infty\left(\hat{\Pi}_z\right)\right]_{11}^{-1}} \frac{\mu_{k_z}}{n}}.$$

We replace $\mathrm{Var}_\infty(\hat{\Pi}_z)$ in Equation (13) with $\frac{1}{n} \mathbf{Z}' \mathrm{M}_{\mathbf{X}} \Psi \mathrm{M}_{\mathbf{X}} \mathbf{Z}$, where $\Psi = \mathrm{diag}(\{\Psi_g\}_{g=1}^G)$, with $\Psi_g = \phi\, \boldsymbol{\iota}_g \boldsymbol{\iota}_g' + (1 - \phi) \left[\mathrm{diag}(\{f(\mathbf{z}_{1,g}, \kappa)\}_{g=1}^G)\right]^{\circ 2}$, and $A^{\circ 2}$ is the Hadamard power of matrix A.

When $\kappa = 0$ and $\eta = 0$, or cluster random effects with the same number of observations per cluster, Nagar's (1959) approximation for the bias of the IV estimator is

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}} - \theta \middle| \mathbf{W}\right] \approx (k_z - 2) \left(n c_z^2\right)^{-1} \left(\phi \bar{n} (1 - \lambda) \rho + (1 - \phi) \varrho\right),^{15} \tag{14}$$

where $\bar{n} = n/G$. Equation (14) makes explicit that the within-cluster correlation ρ counts toward the bias of the IV estimator $\bar{n}$ times more than the correlation of the idiosyncratic term ϱ does. By

[15] See the derivation in Section S.6 of the Supplement.

setting $\rho = \varrho$, the approximate bias of the IV estimator can be rewritten as

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}} - \theta \middle| \mathbf{W}\right] \approx (\mu_{k_z})^{-1} \left[\frac{(k_z - 2)}{k_z}\rho\right]. \tag{15}$$

This result is similar to the bias for $\hat{\theta}_{\mathrm{IV}}$ derived in Bun and de Haan (2010) and Olea and Pflueger (2013).

6.2 Simulation results

Our results are based on $10,000$ Monte Carlo experiments with 199 and 999 bootstrap replications for size and power results, respectively. In repeated Monte Carlo experiments, a small number of bootstrap simulations is sufficient because the bootstrap sampling error cancels out across the Monte Carlo replications. In practice, however, a higher number of bootstrap replications should be used (Davidson and MacKinnon, 2000). We set a 5% significance level for calculating the rejection rates; set $\lambda = 0.01$, which gives small variation on the instruments; and set $\phi = 0.5$, which indicates a design with considerable intra-cluster correlation. [16]

We choose κ, the parameter that sets the degree of variance-preserving heterogeneity in the errors, from the set $\{0, 1, 2\}$, which indicate no heterogeneity (cluster random effects), heterogeneity, and strong heterogeneity. The number of observations in each cluster is controlled by η, which is selected from $\{0, 1, 2\}$. For example, when 400 observations ($n = 400$) are allocated into 20 clusters ($G = 20$), $\eta = 0$ means 20 observations per cluster, $\eta = 1$ means the observations per cluster vary in the range of 12 to 33, and $\eta = 2$ means that this range becomes 6 to 47. We investigate cases with $\rho = \varrho = 0.20$, 0.70, and 0.95, which signify a low to high degree of endogeneity.

We study cases where the noncentrality parameter μ_{k_z} takes values of 0.1 and 18, indicating very weak and strong instruments, respectively. In our baseline model ($\kappa = 0$ and $\eta = 0$) and using Gaussian approximations, the ratio $\left[\frac{(k_z-2)}{k_z}\rho\right] \times (BM)^{-1}$ is asymptotically bounded by one, where BM is the benchmark indicating the worst case of IV estimator bias (see Olea and Pflueger Theorem 1). This property implies that the fraction of the IV bias relative to the benchmark is approximately 5.5% at $\mu_{k_z} = 18$, which is below the 10% tolerance level for setting the 5% critical value of the effective F-test. In our experiments, the rejection rates of the effective F-tests remain above 95% even for the cases where $\kappa \neq 0$ and $\eta \neq 0$, with the estimated IV bias below 7% in all cases and below 4% in the immense majority of cases.

[16] Simulations for different values of λ and $\phi < 0.7$ give similar results. For values of $\phi > 0.7$, the effect of heteroskedasticity becomes smaller because the share of error variance due to the idiosyncratic error $\boldsymbol{\psi}_{1,g} \odot f\left(\mathbf{z}_{1,g}, \kappa\right)$ decreases.

Due to space constraints, we are only reporting a fraction of the Monte Carlo experiments. Further results in the Supplement includes different combinations of μ_{k_z} and ρ for the Wald, KLM, and CLR tests and their bootstrap counterparts. In the Supplement, we also report results for the EE bootstrap tests with multinomial weights, the classical Wald pairs bootstrap, and Davidson and MacKinnon's bootstraps adapted to the cluster case.[17]

6.2.1 Size results

We set $\theta = 0$ as the true value under the DGP. Similar to Hausman and Palmer (2012) and MacKinnon (2013), the instruments $\{\mathbf{z}_g\}_{g=1}^{G}$ are obtained from $\{\mathbf{d}_g\}_{g=1}^{G}$ and $\{\boldsymbol{\vartheta}_g\}_{g=1}^{G}$ sampled from independent log-normal distributions because doing so generates some extreme observations. These extreme values increase with the sample size, making inference under the cluster-robust set-up very difficult. We report the rejection rates for the asymptotic Wald and AR tests together with the Wald multi-equation efficient (ME-eff) bootstrap, the AR estimation equation (EE), and the single equation inefficient (SE–in) and efficient (SE–eff) bootstraps. We also include the rejection rates for the effective F-test with the desired threshold of 10%.[18]

We first investigate the performance of the asymptotic and bootstrap tests assuming different DGPs for the errors $\boldsymbol{\varepsilon}_g$ and $\boldsymbol{\psi}_g$, and noncentrality parameter set at $\mu_{k_z} = 18$. In Table 3, Panels A, B, and C refer to errors sampled from a standard normal distribution, a chi-squared distribution with two degrees of freedom, and a Student t distribution with four degrees of freedom. The errors sampled from the chi-squared and Student t distributions are standardized to have zero mean and unit variance, guaranteeing that the overall variance is the same across panels. The sample size is 400 observations allocated in 20 clusters ($G = 20$), and the number of excluded instruments k_z is 5.

[Table 3 about here.]

The rejection rates of the asymptotic Wald test differ considerably from the nominal level of 5% in all panels, and this difference is increasing with respect to the degree of residual heterogeneity (κ). The asymptotic AR test is also oversized, with rejection rates varying along with the heterogeneity in the number of observations across clusters (η). Interestingly, when comparing both asymptotic tests, the AR size distortion is lower when $\eta = 0$ and $\eta = 1$, but the Wald test outperforms the AR test under normal and *t* errors when $\eta = 2$.

[17] Wald bootstrap tests not reported in this manuscript overreject the true null assumption in almost all cases and have inferior performance in terms of size when compared to the Wald ME-eff bootstrap.

[18] The conservative version of the effective F-test gives rejection rates slightly smaller than those of the standard one. The results are in the Supplement together with the first-stage F-test.

The proposed Wald ME-eff and AR bootstrap tests have rejection rates near the nominal level in all scenarios, with the AR bootstrap test closer to 5% than the Wald ME-eff test in the majority of the experiments. For the AR bootstrap methods, the EE bootstrap rejection rates are smaller than the SE bootstrap rejection rates across panels, κ, and η.

In general, for a given bootstrap procedure, the Rademacher (R) weights deliver better results than the gamma weights when the errors are sampled from standard normal and Student t distributions for a given bootstrap procedure, as expected. Interestingly, however, when the errors are sampled from the standardized chi-squared distribution, which is clearly not symmetric, gamma weights do not necessarily outperform Rademacher (R) weights. Additionally, under Rademacher weights, the AR SE-eff bootstrap rejection rates are closer to the nominal level than the remaining AR bootstrap methods in almost all cases.

We next investigate the performance of the bootstrap tests when the number of exogenous instruments increases but the sample size is kept constant. As shown in Andrews and Stock (2007), the AR test does not have the correct asymptotic size if the magnitude of $\frac{k_z^{\frac{3}{2}}}{n} > 0$ as $n \to +\infty$. In this experiment, the number of clusters G and the sample size n are again set at 20 and 400, respectively, assuming the same number of observatios across clusters ($\eta = 0$). We consider DGPs with noncentrality parameter set at $\mu_{k_z} = 18$ and 0.1, and standard normal errors. The results are reported in Table 4. The columns vary with the κ parameter, which is the degree of variance-preserving heterogeneity in the errors. The number of instruments k_z in the third row increases from 2 to 15. Because the AR test is invariant with respect to μ_{k_z}, we only report it once on the bottom of the table.

[Table 4 about here.]

The results show that the rejection rates of the asymptotic Wald and AR tests are well above the nominal size of 5%. In particular, the asymptotic AR test increases along with the number of instruments, reaching values above 90% at $k_z = 15$. The AR bootstrap tests, on the other hand, have rejection rates closer to the nominal size in all scenarios, including the scenarios with 15 instruments. There is no AR bootstrap test that outperforms the others in all cases. We note, however, that, with Rademacher weights, the AR SE bootstrap tests have rejection rates closer to the nominal level than the AR EE bootstrap test has.

Only when the instruments are strong ($\mu_{k_z} = 18$) does the Wald ME-eff bootstrap test have rejection rates closer to the nominal level when compared with the asymptotic statistic, but the AR bootstrap test rejection rates are closer still. With weak instruments, the Wald ME-eff bootstrap is sized distorted in all scenarios, with overrejection increasing with the number of instruments.

We now explore the impact of endogeneity, which is set by ρ, the correlation between first and second stage residuals. We use a sample design similar to the previous case with five instruments. The results are in Table 5.

[Table 5 about here.]

The asymptotic Wald tests overreject the null hypothesis for all combinations of μ_{k_z} and ρ. We note, however, that only when $\mu_{k_z} = 0.1$ do the rejection rates of the Wald test increase with ρ . This result is consistent with the weak IV literature (see Kleibergen (2002) and Davidson and MacKinnon (2010)). The reason is that the Wald test depends on the IV estimator $\hat{\theta}_{\text{IV}}$, whose approximate bias is given by Equation (15). From this equation, we observe that the effect of ρ on the bias is larger when μ_{k_z} is small, and the opposite when μ_{k_z} is large.

The AR test rejection rates are repeated across columns for a given degree of heterogeneity κ because the AR test is invariant with respect to ρ. Table 5 also shows that the AR bootstrap rejection rates are closer to the nominal size of the tests as compared with the Wald ME-eff, and the AR SE-eff bootstrap with Rademacher weights outperforms the other bootstraps.

In the final size experiment, we study the behavior of the tests when the number of clusters increases. Table 6 shows the rejection rates of asymptotic and bootstrap tests when the number of clusters G increases from 10 to 80 (from 200 to 1600 observations, respectively). We again consider the case with normally distributed errors and five instruments. The noncentrality parameter μ_{k_z} is set at 18 and 0.1 and is kept constant across G. The panels vary in η, and the columns vary in the degree of error heterogeneity κ.

[Table 6 about here.]

The rejections rates of both the asymptotic Wald with strong instruments and AR tests approach the nominal size as the number of clusters G increases; however, the asymptotic Wald test still overrejects the null when the number of clusters is $G = 80$. Nevertheless, in all panels and for all values of G, the proposed AR bootstraps have rejection rates close to the nominal size irrespective of the degree of error heteroskedasticity κ and the number of clusters. Only when instruments are strong does the Wald ME-eff bootstrap tests have rejection rates close to the nominal values, but they are not as close as compared with the AR bootstrap rejection rates. In the case of weak instruments, the Wald and its bootstrap tests always overreject the null.

The previous results show that the proposed AR bootstrap method performs remarkably well across all different specifications, with the Rademacher (R) weights outperforming the gamma weights (Γ) in terms of size in the majority of cases.

6.2.2 Power comparison

Given that the proposed AR bootstrap methods have similar performance in terms of size, we now examine how they vary in terms of power. As opposed to the previous subsection, the instruments for this analysis $\{\mathbf{z}_g\}_{g=1}^G$ are obtained from $\{\mathbf{d}_g\}_{g=1}^G$ and $\{\boldsymbol{\vartheta}_g\}_{g=1}^G$ sampled from independent normal distributions.[19] To save space, we report only results for the two AR bootstrap tests, the EE and SE-eff, with some heteroskedasticity in the errors ($\kappa = 1$), Rademacher bootstrap weights, and errors sampled from standard normal distributions. Results with different degrees of heterogeneity ($\kappa = 0$ and $\kappa = 2$) are similar to those reported below and are included in the Supplement as well. The power curves of the AR SE-in bootstrap test are close to those obtained from the AR SE-eff bootstrap test and are power-dominated when $\eta = 2$, so these results are also not reported.

In the right column in Figure 2, the endogeneity is low ($\rho = 0.20$), while in the left column, endogeneity is high ($\rho = 0.95$). The rows vary in η, the parameter capturing the heterogeneity in the number of observations within clusters. The x-axis represents the value of θ under the true DGP. The curves are obtained by testing the null assumption $H_0 : \theta = 0$ against the alternative $H_1 : \theta \neq 0$ at the 5% significance level.

[Figure 2 about here.]

The graphs show that the asymptotic AR test has rejection rates above 20% when the null assumption $H_0 : \theta = 0$ is true and the nominal level is 5%, whereas the bootstrap test has rejection rates very close to 5% at the same point.

In our experiments, the simulation results show that the SE-eff bootstrap method power dominates the EE bootstrap when $\eta = 0$ and $\eta = 1$, whereas with strong differences in observations across clusters ($\eta = 2$), the opposite result is observed. Therefore, there is no dominant AR bootstrap test in terms of power.

Overall, the performance of the AR SE-eff bootstrap test in terms of size is at least as good as those of the other bootstrap methods. The AR SE-eff bootstrap test also has better power

[19] The results for instruments sampled from a lognormal distribution are similar to the those based on a normal distribution; however, the power curves based on lognormal instruments are flatter compared to those obtained from normally distributed instruments.

properties compared with the AR EE bootstrap test when heterogeneity and cluster size are not extreme.

7 Conclusion

Inference in a linear IV model with clustered errors can be very misleading using asymptotic methods when the number of clusters is small, even when the instruments are strong. We propose two bootstrap methods for the cluster-robust version of the AR test, the estimating equations and the single-equation residual bootstraps. These methods impose the null assumption for generating the bootstrap scores and residuals and are akin to wild bootstrap methods.

In two empirical applications—the study of how institutions affect economic growth and the impact of the economy on civil conflict—we demonstrated the importance of bootstrapping for estimating confidence sets. In the first application, the AR bootstrap confidence regions are much larger than the AR asymptotic confidence region when the instruments are weak and are smaller when the instruments are strong (measured by the effective F-test). In the second application, both the Wald and AR asymptotic and bootstrap confidence regions are also very different from the asymptotic ones. In particular, the AR asymptotic and bootstrap confidence regions indicate no evidence that poor economic performance affects the probability of a civil conflict arising in sub-Saharan countries.

Our simulations show that asymptotic tests are size-distorted when the instruments are strong and the number of clusters is relatively large. The same simulations show that our proposed bootstrap methods for the AR test present rejection probabilities close to the nominal size. These results are robust to different error distributions, the degree of heterogeneity, the numbers of observations per cluster, and the number of exogenous instruments. In terms of power, the single-equation residual bootstrap dominates the score bootstrap in most cases; however, the score bootstrap dominates the residual bootstrap when the difference in observations across clusters is considerable.

We have found that asymptotic tests can have the incorrect size with as many as 80 clusters and with first stages considerably stronger than those considered appropriate in the empirical literature. At a minimum, our wild bootstrap techniques could be broadly used as diagnostic tools and to conduct inference, especially when there is any uncertainty about the impact of cluster sampling or instrument weakness. The empirical applications in the paper also demonstrate that bootstrapping is consequential for inference in specifications that add controls, stratify the sample, reduce sample variation, or reduce the number of clusters. In empirical work, these may be cases that are not the main specification, but are used as robustness checks. The wild bootstrap becomes even more useful under those situations.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review* 91(5): 1369–1401.

———. 2012. The Colonial Origins of Comparative Development: An Empirical Investigation: Reply. *American Economic Review* 102(6): 3077–3110.

Albouy, David Y. 2012. The Colonial Origins of Comparative Development: An Empirical Investigation: Comment. *American Economic Review* 102(6): 3059–3076.

Anderson, Theodore W. and Herman Rubin. 1949. Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *Annals of Mathematical Statistics* 20(1): 46–63.

Andrews, Donald W.K. and James H. Stock. 2007. Testing with many weak instruments. *Journal of Econometrics* 138(1): 24 – 46. 50th Anniversary Econometric Institute.

Andrews, Isaiah. 2016. Conditional Linear Combination Tests for Weakly Identified Models. *Econometrica* 84(6): 2155–2182.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2018. Weak Instruments in IV Regression: Theory and Practice.

Arellano, Manuel. 1987. Computing Robust Standard Errors for Within-groups Estimators. *Oxford Bulletin of Economics and Statistics* 49(4): 431–434.

Bun, Maurice and Monique de Haan. 2010. Weak Instruments and the First Stage F-Statistic in IV Models With a Nonscalar Error Covariance Structure. UvA Econometrics Discussion Paper 2010/02.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics* 90(3): 414–27.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald. 2016. Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity. *Review of Economics and Statistics* .

Chernozhukov, Victor and Christian Hansen. 2008. The Reduced Form: A Simple Approach to Inference with Weak Instruments. *Economics Letters* 100(1): 68–71.

Davidson, Russell and James G. MacKinnon. 2000. Bootstrap Tests: How Many Bootstraps? *Econometric Reviews* 19(1): 55–68.

———. 2008. Bootstrap Inference in a Linear Equation Estimated by Instrumental Variables. *Econometrics Journal* 11(3): 443–77.

———. 2010. Wild Bootstrap Tests for IV Regression. *Journal of Business and Economic Statistics* 28(1): 128–44.

Finlay, Keith and Leandro M. Magnusson. 2009. Implementing Weak Instrument Robust Tests for a General Class of Instrumental Variables Models. *Stata Journal* 9(3): 398–421.

Gelbach, Jonah B., Jonathan Klick, and Thomas Stratmann. 2007. Cheap Donuts and Expensive Broccoli: The Effect of Relative Prices on Obesity. Working paper.

Hausman, Jerry and Christopher Palmer. 2012. Heteroskedasticity-robust inference in finite samples. *Economics Letters* 116(2): 232 – 235.

Hu, Feifang and John D. Kalbfleisch. 2000. The Estimating Function Bootstrap. *Canadian Journal of Statistics* 28(3): 449–81.

Hu, Feifang and James V. Zidek. 1995. A Bootstrap Based on the Estimating Equations of the Linear Model. *Biometrika* 82(2): 263–75.

Imbens, Guido W. and Michal Kolesár. 2016. Robust Standard Errors in Small Samples: Some Practical Advice. *The Review of Economics and Statistics* 98(4): 701–712.

Kleibergen, Frank. 2002. Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression. *Econometrica* 70(5): 1781–1803.

———. 2011. Improved Accuracy of Weak Instrument Robust GMM Statistics through Bootstrap and Edgeworth Approximations. Unpublished manuscript.

Kleibergen, Frank and Richard Paap. 2006. Generalized Reduced Rank Tests Using the Singular Value Decomposition. *Journal of Econometrics* 133(1): 97–126.

Kline, Patrick and Andres Santos. 2012. A Score Based Approach to Wild Bootstrap Inference. *Journal of Econometric Methods* 1(1): 23–41.

Liu, Regina Y. 1988. Bootstrap Procedures under Some Non-I.I.D. Models. *Annals of Statistics* 16(4): 1696–1708.

MacKinnon, James G. 2013. *Thirty Years of Heteroskedasticity-Robust Inference*, pp. 437–61. New York, NY: Springer New York.

MacKinnon, James G. and Matthew D Webb. 2014. Wild Bootstrap Inference for Wild Different Cluster Sizes. Queen's Economics Deparment Working Paper No. 1314.

MacKinnon, James G. and Matthew D. Webb. 2016. Randomization Inference for Difference-in-Differences with Few Treated Clusters. Queen's Economics Department Working Paper No. 1355.

Mammen, Enno. 1993. Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *Annals of Statistics* 21(1): 255–285.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. Economic Shocks and Civil Conflict: An Instrumental Variables Approach. *Journal of Political Economy* 112(4): 725–53.

Moreira, Marcelo J. 2003. A Conditional Likelihood Ratio Test for Structural Models. *Econometrica* 71(4): 1027–48.

Moreira, Marcelo J., Jack R. Porter, and Gustavo A. Suarez. 2009. Bootstrap Validity for the Score Test when Instruments May Be Weak. *Journal of Econometrics* 149(1): 52–64.

Moulton, Brent R. 1990. An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units. *Review of Economics and Statistics* 72(2): 334–38.

Nagar, A. L. 1959. The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica* 27(4): 575–95.

Olea, José Luis Montiel and Carolin Pflueger. 2013. A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics* 31(3): 358–369.

Sanderson, Eleanor and Frank Windmeijer. 2016. A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics* 190(2): 212 – 221.

Staiger, Douglas and James H. Stock. 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica* 65(3): 557–86.

Stock, James H. and Motohiro Yogo. 2005. Testing for Weak Instruments in Linear IV Regression. In Donald W.K. Andrews and James H. Stock, editors, *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge: Cambridge University Press.

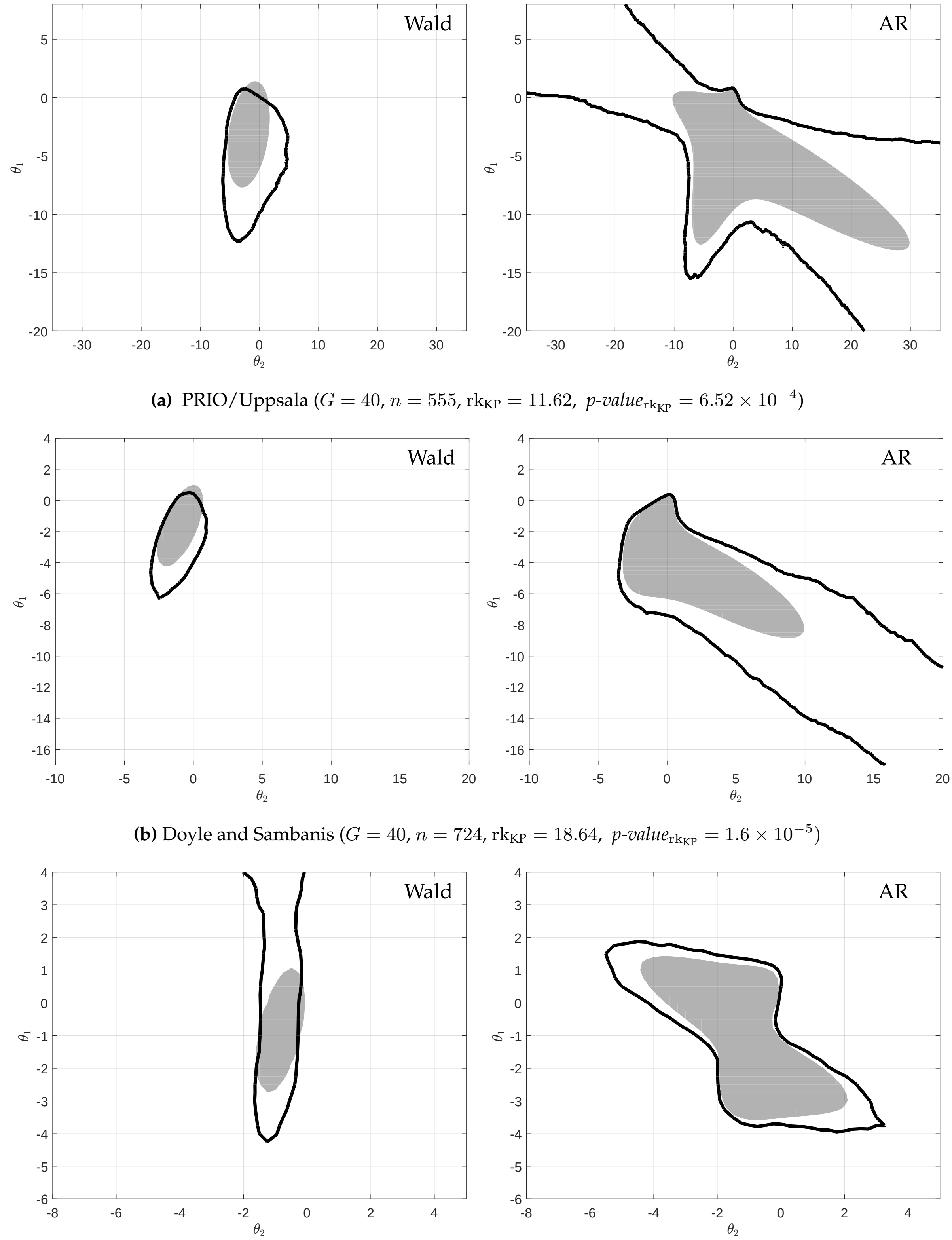
White, Halbert. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 48(4): 817–38.

Wu, C. F. Jeff. 1986. Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis. *Annals of Statistics* 14(4): 1261–95.

Young, Alwyn. 2016. Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covaraince Matrices using Effective Degrees of Freedom Corrections.

Zhan, Zhaoguo. 2014. Detecting Weak Identification by Bootstrap. School of Economics and Management, Tsinghua University, China, Mimeo.

Figure 1: Asymptotic and bootstrapped 95% Wald and AR confidence sets from a replication of Miguel et al. (2004)



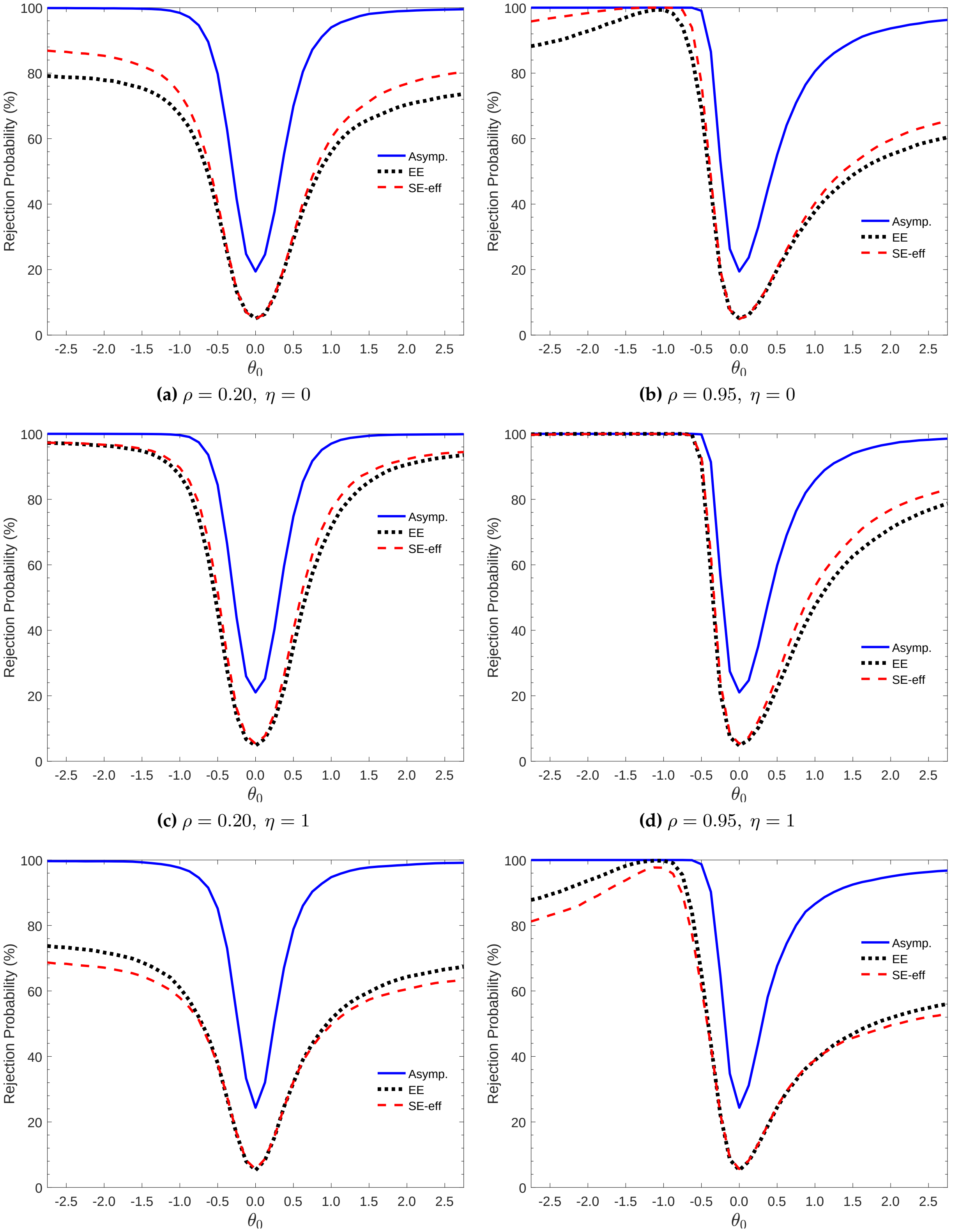
(a) PRIO/Uppsala ($G = 40, n = 555$, $\text{rk}_{\text{KP}} = 11.62$, *p-value*$_{\text{rk}_{\text{KP}}} = 6.52 \times 10^{-4}$)

(b) Doyle and Sambanis ($G = 40, n = 724$, $\text{rk}_{\text{KP}} = 18.64$, *p-value*$_{\text{rk}_{\text{KP}}} = 1.6 \times 10^{-5}$)

(c) Fearon and Laitin ($G = 41$, $n = 743$, $\text{rk}_{\text{KP}} = 18.69$, *p-value*$_{\text{rk}_{\text{KP}}} = 1.5 \times 10^{-5}$)

Notes: In all specifications, θ_1 is GDP growth, θ_2 is lagged GDP growth, the included instruments are country-specific fixed effects and time trends, and the excluded instruments are growth in rainfall and lagged growth in rainfall. The first, second, and third rows correspond to specifications in Table 6, Column 1; Table C3, Column 4; and Table C3, Column 5 from the original paper. The test statistics in the captions are asymptotic. rk_{KP} is the Kleibergen and Paap (2006) rank statistic.

Figure 2: Power comparisons of asymptotic and bootstrapped AR statistics, 10,000 Monte Carlo simulations



(a) $\rho = 0.20,\ \eta = 0$

(b) $\rho = 0.95,\ \eta = 0$

(c) $\rho = 0.20,\ \eta = 1$

(d) $\rho = 0.95,\ \eta = 1$

(e) $\rho = 0.20,\ \eta = 2$

(f) $\rho = 0.95,\ \eta = 2$

Notes: Authors' calculations with 999 bootstrap replications for each simulation using Rademacher weights and a DGP with normal errors. The sample size is 400 observations with 20 clusters: $\eta = 0$ indicates 20 observations per cluster, and $\eta = 1$ and $\eta = 2$ indicate observations per cluster in the range of 12 to 29 and 7 to 42, respectively. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Noncentrality parameter of first-stage F-test: $\mu_{k_z} = 18$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f\left(\mathbf{z}_{1,g}, \kappa\right) = h\left(\kappa\right)\left(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g}\right)^{\kappa}$ evaluated at $\kappa = 1$.

Table 1: Asymptotic and bootstrapped confidence intervals from a replication of Acemoglu et al. (2012)

Specification		Original AJR series capped at 250	ARJ mortality series, capped at 250; Albouy campaign dummy	ARJ mortality series, capped at 250; minimal correction to Albouy campaign dummy	ARJ mortality series, capped at 250; extended correction to Albouy campaign dummy
		Tab. 1, Col. 2	Tab. 3, Col. 2	Tab. 3, Col. 4	Tab. 3, Col. 6
Statistic	Method	(1)	(2)	(3)	(4)
No covariates, ($G = 36, n = 64$)					
Wald	Asymp.	[0.55, 1.08]	[0.43, 1.30]	[0.48, 1.19]	[0.53, 1.08]
	ME-eff.	[0.56, 1.21]	[0.47, 1.53]	[0.52, 1.39]	[0.55, 1.22]
AR	Asymp.	[0.61, 1.46]	[0.56, 2.04]	[0.59, 1.94]	[0.59, 1.51]
	SE-eff.	[0.61, 1.48]	[0.53, 2.24]	[0.58, 1.80]	[0.59, 1.53]
F_{eff}	Asymp.	28.1	13.8	19.2	26.3
With latitude, ($G = 36, n = 64$)					
Wald	Asymp.	[0.51, 1.08]	[0.34, 1.35]	[0.41, 1.22]	[0.50, 1.08]
	ME-eff.	[0.48, 1.24]	[0.35, 1.74]	[0.42, 1.46]	[0.48, 1.26]
AR	Asymp.	[0.54, 1.67]	[0.47, 2.91]	[0.50, 2.06]	[0.53, 1.65]
	SE-eff.	[0.50, 1.61]	[0.16, 3.65]	[0.40, 2.44]	[0.51, 1.68]
F_{eff}	Asymp.	19.3	9.4	13.7	19.7
Without neo-Europes, ($G = 33, n = 60$)					
Wald	Asymp.	[0.50, 1.58]	[0.36, 1.90]	[0.44, 1.79]	[0.48, 1.59]
	ME-eff.	[0.52, 2.02]	[0.46, 2.86]	[0.49, 2.38]	[0.52, 2.16]
AR	Asymp.	[0.65, 2.95]	[0.60, 3.77]	[0.61, 3.22]	[0.62, 3.35]
	SE-eff.	[0.64, 2.89]	[0.52, 5.64]	[0.55, 4.32]	[0.60, 3.26]
F_{eff}	Asymp.	11.3	6.9	8.6	10.3
Without Africa ($G = 19, n = 37$)					
Wald	Asymp.	[0.41, 0.80]	[0.39, 0.93]	[0.40, 0.87]	[0.41, 0.81]
	ME-eff.	[0.40, 0.92]	[0.43, 1.28]	[0.40, 1.13]	[0.41, 0.92]
AR	Asymp.	[0.39, 1.03]	[0.46, 7.66]	[0.43, 2.08]	[0.40, 1.06]
	SE-eff.	[0.42, 1.03]	[0.48, 2.90]	[0.46, 1.71]	[0.44, 1.25]
F_{eff}	Asymp.	46.0	23.0	37.5	51.1
With continent dummies ($G = 36, n = 64$)					
Wald	Asymp.	[0.38, 1.19]	[0.32, 1.31]	[0.35, 1.29]	[0.38, 1.20]
	ME-eff.	[0.35, 1.36]	[0.33, 1.66]	[0.34, 1.54]	[0.35, 1.41]
AR	Asymp.	[0.38, 1.61]	[0.37, 2.25]	[0.37, 1.76]	[0.38, 1.62]
	SE-eff.	[0.27, 1.64]	[0.24, 2.54]	[0.24, 1.92]	[0.26, 1.73]
F_{eff}	Asymp.	10.6	6.8	8.6	9.9
With continent dummies and latitude ($G = 36, n = 64$)					
Wald	Asymp.	[0.33, 1.28]	[0.25, 1.42]	[0.28, 1.40]	[0.33, 1.27]
	ME-eff.	[0.24, 1.40]	[-0.05, 2.00]	[0.17, 1.68]	[0.23, 1.36]
AR	Asymp.	[0.23, 1.66]	[0.21, 3.50]	[0.26, 2.03]	[0.24, 1.56]
	SE-eff.	[-0.05, 1.77]	[-0.25, 8.64]	[-0.21, 2.43]	[-0.04, 1.68]
F_{eff}	Asymp.	7.7	4.8	6.1	7.6
With percent of European descent in 1975 ($G = 36, n = 64$)					
Wald	Asymp.	[0.34, 1.07]	[0.13, 1.34]	[0.23, 1.21]	[0.32, 1.07]
	ME-eff.	[0.28, 1.21]	[0.14, 1.84]	[0.06, 1.43]	[0.21, 1.27]
AR	Asymp.	[0.18, 1.45]	[0.14, 3.69]	[0.05, 1.75]	[0.09, 1.37]
	SE-eff.	[0.13, 1.49]	[-0.10, 5.34]	[-0.03, 2.27]	[0.03, 1.47]
F_{eff}	Asymp.	12.9	6.2	9.9	12.8
With malaria, ($G = 35, n = 62$)					
Wald	Asymp.	[0.24, 0.80]	[0.00, 0.96]	[0.16, 0.88]	[0.25, 0.80]
	ME-eff.	[0.23, 0.95]	[-0.78, 1.17]	[-0.05, 1.09]	[0.23, 0.94]
AR	Asymp.	[0.06, 1.15]	$(-\infty, +\infty)$	$(-\infty, +\infty)$	[0.05, 1.15]
	SE-eff.	[-0.61, 1.40]	$(-\infty, +\infty)$	[-2.44, -2.42] ∪ [-2.39, 2.63]	[-0.84,-0.80] ∪ [-0.66, 1.31]
F_{eff}	Asymp.	11.5	5.8	8.5	12.0

Notes: The wild bootstraps use Rademacher resampling weights. The critical values of the effective F-test are 23.1 (5%) and 19.7 (10%) under a 10% tolerance for bias, and 15.1 (5%) and 12.4 (10%) under a 20% tolerance for bias. This test sets the ratio of the Nagar bias of the TSLS over the worst-case-scenario bias. If this ratio is above the threshold, then the instruments are weak. As the bias tolerance increases, lower the critical value of the effective F-test is.

Table 2: Summary of p-values ($\times 100$) for $H_0 : \theta_0 = 0$ from a replication of Acemoglu et al. (2012)

	Original AJR series		ARJ mortality series, Albouy campaign dummy		ARJ mortality series, minimum correction to Albouy campaign dummy		ARJ mortality series, extended correction to Albouy campaign dummy	
Specification	AR asym.	AR wild boot.	AR asym.	AR wild boot.	AR asym.	AR wild boot.	AR asym.	AR wild boot.
No covariates	0.04	0.03	0.01	0.03	0.06	0.00	0.10	0.03
With latitude	0.60	0.43	0.27	3.07	0.22	1.63	0.55	0.67
Without neo-Europes	0.09	0.03	0.04	0.63	0.14	0.43	0.27	0.40
Without Africa	0.94	2.13	1.27	0.20	1.39	0.53	0.89	1.97
With continent dummies	0.88	1.90	1.04	1.93	0.78	2.07	0.86	2.17
With continent dummies and latitude	2.44	5.87	3.03	7.50	2.01	7.27	2.27	5.84
With percent of European descent in 1975	3.17	2.57	3.03	6.50	4.48	5.70	4.09	4.30
With malaria	4.32	7.97	17.29	22.41	9.31	12.67	4.54	7.60

Notes: The wild bootstraps use Rademacher resampling weights. The lightest shade of gray indicates if the test is not rejected at the 1% significance level, the medium gray at the 5% level, the medium-dark at the 10% level, and the darkest at the 20% level. Mortality rates have been capped at 250.

Table 3: Test rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGPs with different distributions of random errors, 10,000 Monte Carlo simulations

			$\eta = 0$			$\eta = 1$			$\eta = 2$		
		$\kappa \rightarrow$	0	1	2	0	1	2	0	1	2
Stat.	Method	ω_g	Panel A: $\mathcal{N}$ errors								
Wald	Asymp.		33.70	41.88	45.73	24.17	25.38	29.34	24.66	25.72	26.21
	ME-eff	Γ	0.52	0.77	1.19	5.81	6.22	6.75	8.84	9.19	9.15
		R	1.42	2.06	2.40	5.02	5.23	5.55	5.82	5.86	5.99
AR	Asymp.		17.08	16.99	16.95	19.01	19.57	19.54	27.07	26.58	27.05
	EE	Γ	6.30	6.23	6.17	6.55	6.57	6.06	7.59	7.16	7.23
		R	4.46	4.48	4.43	4.32	4.38	4.46	4.20	4.38	4.65
	SE-in	Γ	6.89	6.90	6.82	7.12	7.21	6.66	7.70	7.24	7.41
		R	5.39	5.28	5.24	5.58	5.59	5.17	5.51	5.40	5.42
	SE-eff	Γ	6.58	6.92	6.77	6.87	7.19	6.65	7.82	7.66	7.71
		R	5.07	5.00	5.00	5.18	5.14	4.89	5.51	5.20	5.36
F$_{eff}$			96.57	99.72	99.94	98.95	99.46	99.82	97.82	98.87	98.96
			Panel B: χ^2 errors								
Wald	Asymp.		34.75	41.71	50.10	22.69	24.09	29.12	21.48	24.24	24.78
	ME-eff	Γ	0.08	0.51	0.60	2.69	3.65	4.43	5.97	7.09	7.17
		R	0.53	1.24	1.46	2.45	2.75	3.04	5.07	6.48	6.50
AR	Asymp.		13.33	13.75	13.41	15.93	15.80	15.97	21.85	23.29	23.66
	EE	Γ	4.75	5.24	4.68	5.15	5.44	5.54	5.74	6.44	6.40
		R	3.52	4.34	3.75	3.96	4.08	4.27	3.81	4.43	4.38
	SE-in	Γ	5.00	5.48	5.06	5.48	5.64	5.85	5.84	6.26	6.22
		R	3.87	4.44	3.89	4.31	4.37	4.66	4.45	4.82	4.70
	SE-eff	Γ	5.38	5.72	5.34	5.83	5.75	6.01	6.28	7.03	6.75
		R	3.98	4.41	4.08	4.48	4.44	4.60	4.69	5.12	5.16
F$_{eff}$			92.38	97.71	98.84	96.36	97.47	98.68	94.06	95.99	96.24
			Panel C: t errors								
Wald	Asymp.		31.12	41.02	46.54	22.19	24.81	29.27	22.15	24.36	24.84
	ME-eff	Γ	0.72	1.12	1.64	5.04	5.51	5.95	7.18	7.87	7.83
		R	2.06	2.27	2.97	5.00	4.74	5.32	4.84	5.75	5.77
AR	Asymp.		15.56	15.89	15.22	17.94	17.80	17.64	25.14	26.11	25.99
	EE	Γ	5.51	5.54	5.46	5.75	5.73	5.52	6.68	6.56	6.44
		R	4.02	4.16	4.06	4.14	4.14	4.22	4.29	3.98	4.37
	SE-in	Γ	6.57	6.60	6.29	6.84	6.78	6.59	7.17	7.04	7.23
		R	4.99	5.20	5.04	5.38	5.46	5.34	5.55	5.62	5.82
	SE-eff	Γ	6.63	6.45	6.42	6.79	6.64	6.58	7.50	7.35	7.44
		R	4.67	4.72	4.63	5.10	5.00	4.86	5.21	5.33	5.45
F$_{eff}$			93.43	97.69	98.61	96.52	97.51	98.51	95.54	96.87	96.99

Notes: Authors' calculations with 199 bootstrap replications for each simulation. The sample size is 400 observations with 20 clusters: $\eta = 0$ indicates 20 observations per cluster, and $\eta = 1$ and $\eta = 2$ indicate observations per cluster in the range of 12 to 29 and 7 to 42, respectively. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Degree of endogeneity: $\rho = 0.95$. Skedastic function: $f\left(\mathbf{z}_{1,g}, \kappa\right) = h\left(\kappa\right)\left(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g}\right)^{\kappa}$. The weights Γ and R correspond to the gamma and Rademacher weights, respectively. The effective F-test uses 5% critical values under a 10% tolerance for bias.

Table 4: Test rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGPs with different numbers of instruments (k_z), 10,000 Monte Carlo simulations

			$\kappa = 0$				$\kappa = 1$				$\kappa = 2$			
Stat.	Method	$k_z \rightarrow$	2	5	10	15	2	5	10	15	2	5	10	15
							DGP $\mu_{k_z} = 18$							
Wald	Asymp.		16.16	33.70	24.18	29.50	17.94	41.88	24.24	32.80	21.34	45.73	24.47	36.68
	ME-eff	Γ	7.27	0.52	14.41	3.13	7.44	0.77	14.40	3.81	7.67	1.19	12.91	4.90
		R	5.58	1.42	7.81	5.72	5.70	2.06	8.06	6.00	5.65	2.40	7.66	6.42
F_{eff}			95.01	96.57	99.22	100.00	96.22	99.72	99.64	100.00	97.57	99.94	99.84	100.00
							DGP $\mu_{k_z} = 0.1$							
Wald	Asymp.		69.20	93.67	99.54	99.95	69.93	92.68	99.26	99.89	69.76	91.84	99.29	99.89
	ME-eff	Γ	25.27	44.02	81.01	99.03	26.64	43.14	80.61	98.37	27.29	41.90	79.10	97.88
		R	19.05	31.54	71.45	97.13	20.03	30.57	70.63	95.98	21.34	29.65	69.06	95.09
F_{eff}			2.73	0.29	0.48	20.19	3.12	0.41	0.59	21.84	4.00	0.64	0.61	23.47
							DGP $\mu_{k_z} = 18,\ 0.1$							
AR	Asymp.		6.86	17.08	60.93	96.98	6.59	16.99	61.03	97.03	6.19	16.95	60.44	96.99
	EE	Γ	5.79	6.30	7.37	5.99	5.29	6.23	7.21	6.44	5.15	6.17	7.83	5.86
		R	4.70	4.46	4.30	2.64	4.70	4.48	4.25	2.87	4.70	4.43	4.47	2.53
	SE-in	Γ	5.92	6.89	8.52	7.30	5.81	6.90	8.30	7.64	5.60	6.82	8.66	6.85
		R	5.39	5.39	5.37	5.85	5.30	5.28	5.27	6.00	5.26	5.24	5.47	5.51
	SE-eff	Γ	5.96	6.58	8.78	7.39	5.69	6.92	8.64	7.86	5.64	6.77	8.95	7.09
		R	5.38	5.07	5.02	5.69	5.21	5.00	4.91	5.66	5.04	5.00	5.07	5.36

Notes: Authors' calculations with 199 bootstrap replications for each simulation. The sample size is 400 observations with 20 clusters, 20 observations per cluster. Number of included instruments: $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Degree of endogeneity: $\rho = 0.95$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. The weights Γ and R correspond to the gamma and Rademacher weights, respectively. The effective F-test uses 5% critical values under a 10% tolerance for bias.

Table 5: Test rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGPs with different degrees of endogeneity (ρ), 10,000 Monte Carlo simulations

			$\kappa = 0$			$\kappa = 1$			$\kappa = 2$		
Stat.	Method	$\rho \rightarrow$	0.20	0.70	0.95	0.20	0.70	0.95	0.20	0.70	0.95
						DGP $\mu_{k_z} = 18$					
Wald	Asymp.		33.89	33.85	33.70	42.93	42.46	41.88	47.83	46.60	45.73
	ME-eff	Γ	0.67	0.70	0.52	0.85	0.84	0.77	0.97	1.28	1.19
		R	1.84	1.65	1.42	2.15	2.30	2.06	2.19	2.40	2.40
F_{eff}			96.27	96.50	96.57	99.67	99.64	99.72	99.96	99.91	99.94
						DGP $\mu_{k_z} = 0.1$					
Wald	Asymp.		12.90	58.29	93.67	14.48	60.26	92.68	16.04	60.04	91.84
	ME-eff	Γ	3.37	20.41	44.02	3.51	22.06	43.14	3.78	21.72	41.90
		R	4.25	15.41	31.54	4.28	16.07	30.57	4.48	15.90	29.65
F_{eff}			0.23	0.18	0.29	0.23	0.31	0.41	0.28	0.43	0.64
						DGP $\mu_{k_z} = 18,\ 0.1$					
AR	Asymp.		17.08	17.08	17.08	16.99	16.99	16.99	16.95	16.95	16.95
	EE	Γ	6.15	6.15	6.15	6.25	6.25	6.25	6.12	6.12	6.12
		R	4.32	4.32	4.32	4.14	4.14	4.14	4.25	4.25	4.25
	SE-in	Γ	6.89	6.89	6.89	6.90	6.90	6.90	6.82	6.82	6.82
		R	5.39	5.39	5.39	5.28	5.28	5.28	5.24	5.24	5.24
	SE-eff	Γ	6.58	6.58	6.58	6.92	6.92	6.92	6.77	6.77	6.77
		R	5.07	5.07	5.07	5.00	5.00	5.00	5.00	5.00	5.00

Notes: See notes from Table 4. Degrees of endogeneity (ρ): 0.20, 0.70, 0.95.

Table 6: Test rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGPs with different numbers of clusters (G), 10,000 Monte Carlo simulations

			$\kappa = 0$				$\kappa = 1$				$\kappa = 2$			
Stat.	Method	$G \rightarrow$	10	20	40	80	10	20	40	80	10	20	40	80
			DGP μ_{k_z}=18											
Wald	Asymp.		38.18	33.70	28.08	17.13	43.39	41.88	38.15	24.86	46.77	45.73	44.22	27.98
	ME-eff	Γ	0.21	0.52	1.54	9.09	0.27	0.77	2.87	9.75	0.37	1.19	4.09	10.14
		R	1.52	1.42	2.10	6.11	1.73	2.06	2.91	6.41	2.01	2.40	3.50	6.64
F$_{eff}$			98.83	96.57	94.48	94.22	99.49	99.72	99.88	99.62	99.86	99.94	99.98	99.54
			DGP μ_{k_z}=0.1											
Wald	Asymp.		94.84	93.67	93.25	92.82	94.59	92.68	91.10	91.80	93.80	91.84	89.78	91.12
	ME-eff	Γ	62.27	44.02	32.43	29.79	61.50	43.14	32.64	34.10	57.95	41.90	32.84	35.97
		R	48.64	31.54	22.80	22.44	47.73	30.57	22.61	25.65	45.49	29.65	23.22	27.99
F$_{eff}$			9.50	0.29	0.00	0.00	10.13	0.41	0.04	0.01	11.25	0.64	0.11	0.01
			DGP μ_{k_z}=18, 0.1											
AR	Asymp.		47.50	17.08	8.32	5.63	47.32	16.99	7.70	5.18	46.91	16.95	8.12	4.88
	EE	Γ	6.05	6.30	6.22	6.08	5.87	6.23	5.93	5.86	5.72	6.17	5.92	5.46
		R	2.94	4.46	4.79	5.02	2.97	4.48	4.64	4.95	2.72	4.43	4.65	4.76
	SE-in	Γ	7.62	6.89	6.18	5.95	7.17	6.90	5.90	5.84	7.48	6.82	5.78	5.42
		R	6.65	5.39	4.99	5.11	6.70	5.28	4.70	5.03	6.42	5.24	4.69	4.89
	SE-eff	Γ	7.41	6.58	6.18	5.91	6.81	6.92	6.00	5.89	7.09	6.77	5.89	5.47
		R	5.40	5.07	4.93	5.17	5.22	5.00	4.67	5.20	5.11	5.00	4.68	4.85

Notes: Authors' calculations with 199 bootstrap replications for each simulation. Sample size, number of clusters: $(n, G) = (200, 10)$; $(400, 20)$; $(800, 40)$; $(1600, 80)$. 20 observations per cluster. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Degree of endogeneity: $\rho = 0.95$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. The weights Γ and R correspond to the gamma and Rademacher weights, respectively. The effective F-test uses 5% critical values under a 10% tolerance for bias.